Thematizer:
A Computational Approach to Thematic Theory and Analysis

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität
München

vorgelegt von

Paul Thomas Gahman

aus
Alaska, USA

2024

Referent/in:	Prof. Dr. Christina Sanchez-Stockhammer
Korreferent/in:	Prof. Dr. Gero Kunter
Tag der mündlichen Prüfung:	17.11.2023

Abstract

Thematic theory, comprised of the concepts of theme, rheme, and thematic progression, concerns itself with the interplay between word order, information status, propositional content and discourse function. In contemporary research on thematic theory, researchers have begun to leverage computational means for text analysis with respect to thematic structure. However, deficiencies in both the theoretical treatment and computational operationalization of thematic theory have limited writers' accessibility to thematic structure.

The present work set out to address these deficiencies by identifying remaining gaps in thematic theory and by developing the software Thematizer, which automatically analyzes texts in terms of themes, rhemes and thematic progression. To develop and train Thematizer, 30 Wikipedia articles, L1 and L2 university texts, blog articles and lyrics were used. The accuracy of *Thematizer*, measured with the F_1 score, was then validated with ten novel test texts. All 160 texts were first manually analyzed for comparison against the results that the software yielded. The resulting F_1 scores for Thematizer's parsing functionality were then used as a metric for its operationalization of thematic theory via computational means. In turn, Thematizer's degree of operationalization informed writers' degree of accessibility to thematic theory.

In the identification of themes and rhemes, Thematizer achieved an F_1 score of 85.8% for training texts and 92.0% for test texts (cf. 89.1% gold standard). The identification and classification of marked themes exceeded the gold standard of 89.1% through the training texts' F_1 score of 94.9% and the test texts' F_1 score of 93.4%. Finally, only training texts ($F_1 = 80.2\%$) exceeded the gold standard of 79.2% for the classification of thematic progression patterns, with test texts yielding an accuracy of $F_1 = 75.9\%$.

These findings indicate that Thematizer successfully operationalized marked theme identification and classification but was only able to partially operationalize the identification of themes and rhemes in text. Thematic progression, however, was inconsistently operationalized due to the wide range of F_1 scores that Thematizer achieved and that were often below the gold standard. Operationalization and thereby accessibility to thematic theory were both facilitated by automated means, which represents a marked advancement in the computational treatment of theme, rheme and thematic progression.

Ultimately, the present work was able to forward thematic theory both conceptually and computationally. The inclusion of unmarked themes in conjunction with marked themes enriches thematic analyses by readily tracing GIVEN discourse topics through a text. Further delineation of marked themes into separate types and semantic subclasses reveals their functional, logical and contextualizing contribution to the discourse messages that follow. Visualization of the analytical results from the thematic analyses embedded within the user's text in Thematizer's web interface additionally affords greater interactability with thematic structure. Thematizer's ability to analyze multiple documents and simultaneously present their results facilitates intertextual analyses that previous tools lacked. Including the option to export the results from the thematic analyses also provides users with agency over their own texts for subsequent use in their own research. Finally, the analytical results that Thematizer delivers can enable users to further reflect on the structural and logical development of their text.

Die vorliegende Dissertation befasst sich mit der sogenannten thematischen Theorie, die sich aus den Begriffen Thema, Rhema und Thema-Rhema-Gliederung zusammensetzt. Diese Theorie wurde erstmals im 19. Jahrhundert von Weil (1978, ursprünglich 1887) im Hinblick auf den Zusammenhang zwischen Wortstellung, Informationsgehalt und Diskursentwicklung aufgestellt. Auf der theoretischen Grundlage Weils formalisierte die Prager Schule das Thema-Rhema-Paradigma und trieb es voran. Repräsentative Linguisten dieser theoretischen Richtung waren Mathesius (1983), Firbas (1992) und Daneš (1974). Sie gelten als Vertreter der funktionalen Satzperspektive, die im Gegensatz zu der von Halliday (1967) begründeten systemisch-funktionalen Grammatik steht, die auch von Fries (1995), Hawes (2010a) und Matthiessen (1995) vertreten wird. Gemeinsam ist beiden theoretischen Ansätzen, dass Sprache als funktionales, d.h. kontextbezogenes System fungiert und daher im Hinblick auf kontextuelle Einflüsse betrachtet werden muss, um die zugrunde liegenden grammatikalischen und diskursiven Funktionen jeder Äußerung zu bestimmen. Als Teil dieses Sprachsystems können Äußerungen in Thema und Rhema unterteilt werden.

Diese Gliederung eines Satzes in Thema und Rhema variiert je nach theoretischem Ansatz. Dennoch bilden grundsätzlich die Satzglieder, die nicht zum Prädikat gehören, also das grammatikalische Subjekt und alle vorangestellten Adjunkte, Komplemente und Adverbialien, das Thema. Zum Rhema gehören dagegen alle Satzglieder des Prädikats. Durch die Gliederung der thematischen und rhematischen Elemente eines Textes tritt deren diskursive Funktion zutage: Das Thema fungiert als Grundlage einer kommunikativen Aussage durch wiederholte Realisierung bereits etablierter (d.h. gegebener) Diskursthemen. Gegebene Diskursthemen sind hier in dem Sinne zu verstehen, dass sie zuvor mindestens einmal als explizites Thema oder Rhema im Text realisiert wurden. Demgegenüber trägt das Rhema zur kommunikativen Aussage bei, indem es den Diskurs durch neue Themen weiterentwickelt. Neue Diskursthemen entsprechen nun denjenigen, die im Text erstmals vorgestellt und behandelt werden.

Sobald die Themata und Rhemata eines Textes bestimmt sind, kann ihre Entwicklung anhand der sogenannten Thema-Rhema-Gliederung (Daneš 1974) nachverfolgt werden. Je nachdem, ob dieselben Themata zweier Sätze oder nur das Rhema des vorhergehenden Satzes als Thema des nachfolgenden Satzes realisiert wird, ergeben sich unterschiedliche Thema-Rhema-Schemata. An ihnen lässt sich die sogenannte Methode der Entwicklung im Text (Fries 1995) feststellen, welche die Entwicklung der Diskursthemen und damit den Diskurs selbst durch bewusste Themenwahl konkretisiert. Die Thema-Rhema-Struktur gibt somit Aufschluss darüber, wie der Kontext und Kotext die thematische Struktur eines Textes beeinflussen.

Wie sich die einzelnen Diskursthemen im Laufe des Diskurses als realisierte Themata und Rhemata entwickeln, lässt sich an der Thema-Rhema-Struktur der Textanalyse ablesen. Das folgende Beispiel zeigt, wie ein Text thematisch analysiert werden kann, um seine strukturelle und diskursive Entwicklung aufzuzeigen.

KONSTANTE ENTWICKLUNG (ENGLISCH: *CONSTANT CONTINUOUS PROGRESSION*)

	THEMA	RHEMA
Satz 1	***Wir*** [T_1]	*waren an allen Schritten der Berechnung der Anteile für die Mischung selbst beteiligt..*
Satz 2	***Wir*** [T_2]	*haben auch Prüfverfahren wie den Slump-Test, das Stückgewicht und die Menge der mitgerissenen Luft pro Volumeneinheit verwendet.*

In beiden Sätzen bildet das Pronomen *wir* das grammatikalische Subjekt und damit das Thema und die Basis der kommunikativen Aussage. Die übrigen Satzglieder gehören dementsprechend zum Rhema, das den Diskurs auf der Basis des Themas *wir* entwickelt. Durch die lexikalische Wiederholung des Pronomens als thematischen Subjekts entsteht eine stetige Weiterentwicklung des bereits etablierten Diskursthemas *wir*. Eine solche Struktur lässt auch erkennen, dass das Thema sich mit dem Sachverhalt eines Agens durch *wir* befasst, während das Rhema die angewandte Methodik einer Studie ist. Daraus lässt sich eine parallele Struktur der diskursiven Entwicklung in den beiden Sätzen ableiten, die durch eine Thema-Rhema-Gliederung verdeutlicht wird.

Die Forschung zu Thema und Rhema erfolgte bisher aus textlinguistischer Perspektive. Die Bedeutung dieser Theorie, vor allem in der geschriebenen Sprache, hat sich besonders in der Analyse von Textualität und Textgattung (Fries 1995; Matthiessen 1995; Hawes 2010b), in der Lehre (Lee 2001; Jalilifar 2010; Jingxia & Liu 2013) und in der Übersetzung (McCabe 1999; Jalilifar 2009; Williams 2009) gezeigt. Neuere Studien haben die thematische Theorie vorangetrieben, indem sie die Analyse von Texten hinsichtlich ihrer thematischen Struktur durch computergestützte Ansätze automatisierten (Schwarz et al. 2008; Park & Lu 2015; Domínguez et al. 2020).

Trotz der Weiterentwicklung der thematischen Theorie durch die vorangegangenen Studien traten wesentliche Mängel in den thematischen Modellen und der automatischen Analyse der thematischen Struktur zutage. Erstens wurden in früheren Arbeiten unmarkierte Themata dem Rhema zugeordnet, solange sie zusammen mit markierten Themata realisiert wurden. Es wird argumentiert, dass dies den Beitrag verschleiert, den gegebene Themata zur Entwicklung des Diskurses leisten. Zweitens haben manuelle thematische Analysen zwar die verschiedenen syntaktischen und semantischen Funktionen untersucht, die markierte Themata erfüllen, ihre Typisierung blieb jedoch auf textuelle und interpersonelle markierte Themata allein beschränkt. Darüber hinaus fehlte eine automatisierte semantische Klassifizierung der markierten Themata in der zuvor entwickelten Software. Schließlich wurde festgestellt, dass die Datenvisualisierung, die intertextuelle Analyse und die Bereitstellung der Analyseergebnisse bislang fehlten oder unzureichend waren.

Vor dem Hintergrund der bisherigen Forschung zu Thema und Rhema und der identifizierten Forschungslücken ergaben sich zwei grundlegende Forschungsfragen, denen die vorliegende Dissertation nachgeht. Erstens: Wie können die identifizierten Forschungslücken in der bisherigen Behandlung und Konzeptionalisierung thematischer Theorie überwunden werden? Zweitens: Wie kann der Zugang zur thematischen Theorie für Schreibende unabhängig von ihrem sprachlichen oder linguistischen Hintergrund ermöglicht werden, indem die thematische Analyse operationalisiert und damit durch Software automatisiert wird? Motivation für die erste Forschungsfrage liegt darin, dass die vorliegende Dissertation einen Beitrag zum linguistischen Repertoire der thematischen Theorie leisten soll. Dadurch könnten neue Perspektiven und Theorien zum Thema-Rhema-Paradigma gewonnen werden. Der Grundgedanke hinter der zweiten Forschungsfrage war die Automatisierung der thematischen Analyse im Text, sodass Schreibende einen detaillierten Einblick in den strukturellen und diskursiven Aufbau ihres Textes erhalten.
Zur Beantwortung der beiden Forschungsfragen wurde die Software *Thematizer* entwickelt, die mehrere Texte zerlegen kann und deren Satzglieder automatisch thematisch analysiert. Thematizer wurde in der Programmiersprache Python zusammen mit der **Spacy** API und den Programmbibliotheken **Coreferee** und **Gensim** entwickelt. Diese dienten der programmatischen Analyse der Texte mittels Dependenzanalysen und lexikalischer Implikationsbeziehungen (engl.: *lexical entailment*). Für die Darstellung und

Zusammenfassung der Ergebnisse der thematischen Analysen wurde die **Dash** API verwendet, mit der eine Webseite dynamisch mit Abbildungen generiert werden kann und die als Schnittstelle zwischen den analysierten Daten und der programmatischen Funktionalität von Thematizer dient.

Zur Operationalisierung von Thema, Rhema und deren Gliederung wurden die oben erörterten Prinzipien der thematischen Theorie angewandt. Diese wurden in Thematizer programmiert und definieren die drei grundlegenden Parsing-Analysen, welche die Software bei jeder Textanalyse durchführt. Die erste Parsing-Analyse besteht daraus, die Themata und Rhemata aller Sätze zu identifizieren, nachdem der Text von textuellen Störungen und Unregelmäßigkeiten (d.h. *noise*) bereinigt wurde. Dabei werden die Themata in zwei Klassen eingeteilt: das grammatikalische Thema und die sogenannten markierten Themata (engl.: *marked themes*). Letztere sind für die zweite Parsing-Analyse der Software von Bedeutung, in der sie nach ihrer Zugehörigkeit zu einer markierten Klasse (**circumstantial, structural, modal, hypotactic** und **projecting theme**) und einer semantischen Klasse (z.B. TEMPORAL, LOKATIV oder KONZESSIV) kategorisiert werden. In der dritten und letzten Parsing-Analyse wird der Text hinsichtlich seiner Thema-Rhema-Struktur analysiert. Dazu werden die einzelnen Themata und Rhemata aus der ersten Parsing-Analyse verwendet, um zu bestimmen, welche Thema-Rhema-Schemata zwischen den einzelnen Sätzen bestehen.

Für die Entwicklung der Software und die Überprüfung der produzierten Analysen wurden insgesamt 150 Trainingstexte verwendet, die fünf verschiedenen Texttypen entsprechen: Wikipedia-Artikel, L1- und L2-Aufsätze aus dem universitären Bereich, Blog-Artikel und Songtexte. Insgesamt wurden jeweils 30 Texte pro Texttyp verwendet, die zunächst vom Autor dieser Arbeit manuell thematisch analysiert wurden. Diese Analysen wurden mit den vom Thematizer gelieferten thematischen Daten verglichen, um Änderungen an der programmatischen Funktionalität der Software vorzunehmen. Nach der finalen Entwicklung von Thematizer wurden die Funktionalität und die Ergebnisse der Software anhand von zehn neuen Texten validiert. Die abschließenden Ergebnisse der thematischen Analysen, die Thematizer mittels der Trainings- und Validierungstexte durchführte, wurden mit dem sogenannten F_1-Wert bestimmt, der die Gesamtgenauigkeit der gelieferten Daten beschreibt.

Insgesamt erreichte Thematizer einen F_1-Wert von 85,7% für die Trainingstexte und 85,4% für die Validierungstexte. Damit liegt die Gesamtgenauigkeit von Thematizer mindestens 5,0% über dem Goldstandard von 79,2%, der aus bisherigen Forschungsarbeiten zu computergestützter Textanalyse und thematischer Theorie hervorgeht. Allerdings ist dieses Ergebnis vor dem Hintergrund zu betrachten, dass sich die genannten Gesamtwerte aus den einzelnen F_1-Werten der drei Parsing-Analysen ableiten, die selbst teilweise unter dem Goldstandard lagen. So erzielte die Identifikation thematischer und rhematischer Elemente in den Trainingstexten in der ersten Parsing-Analyse einen F_1-Wert von 85,8% (vgl. Goldstandard: 89,1%). Darüber hinaus erreichte die Thema-Rhema-Gliederung der Validierungstexte mit einem F_1-Wert von 75,9% den niedrigsten F1-Wert aller Parsing-Analysen (vgl. Goldstandard: 79,2%). Alle anderen Parsing-Analysen der Trainings- als auch der Validierungstexte übertrafen jedoch den Goldstandard. Insbesondere die Identifikation und Klassifikation von markierten Themata erwies sich als die genaueste Parsing-Analyse mit der größten Differenz zum Goldstandard: 94,9% (Trainingstexte) und 93,4% (Validierungstexte) im Vergleich zu 89,1% (Goldstandard).

Die verschiedenen Fehler und Fehlerklassen, die zu einer Senkung des F_1-Wertes führten, variierten je nach Parsing-Analyse. Bei der Identifikation von Themata und Rhemata basierten falsche Ergebnisse auf fehlerhaften Dependenzanalysen und Mustervergleichen. Bei der

Identifikation und Klassifikation von markierten Themata wurden T-Units (kurz: zu analysierende Texteinheiten, die meist in syntaktisch unabhängiger Satzform auftreten) falsch zerlegt oder übersehen, sodass neue, fehlerhafte Themata in den Text eingeführt oder bereits vorhandene Themata gänzlich übersehen wurden. Darüber hinaus scheiterte die Bestimmung der indexikalischen Spannweite bei markierten Themata gelegentlich, wenn sogenannte *right dependents*, d.h. abhängige, untergeordnete Satzglieder, auf die falschen Endglieder der markierten Themata verwiesen. Die meisten Fehler traten schließlich bei der Thema-Rhema-Gliederung auf. Dazu gehören die fehlerhafte Auflösung von Implikationsbeziehungen, von Kohäsion durch Koreferenz und lexikalische Wiederholung sowie von thematischen Sprüngen (engl.: *thematic breaks*). Implikationsrelationen, z.B. durch Hyponymie, Meronymie, Synonymie, erwiesen sich als die mit Abstand häufigste Ursache von Parsing-Fehlern bei der Thema-Rhema-Gliederung und erklären damit den besonders niedrigen F_1-Wert bei dieser Aufgabe.

Diese Ergebnisse zeigen, dass die Operationalisierung und damit die Zugänglichkeit der thematischen Theorie nicht durchgängig gelungen ist. Bei der Identifikation von Themata und Rhemata erreichten nur die Validierungstexte den Goldstandard, sodass die Operationalisierung als teilweise gelungen betrachtet werden kann. Die Identifikation und Klassifikation von markierten Themata erwiesen sich als die einzige Parsing-Analyse, die in Training und Validierung den Goldstandard erreichte und damit als erfolgreich operationalisiert gelten kann. Als nicht ausreichend operationalisierbar erwies sich schließlich die Thema-Rhema-Gliederung, da der Goldstandard nicht durchgehend erreicht werden konnte. Die hohe Häufigkeit von Fehlanalysen stellt in diesem Zusammenhang die Zuverlässigkeit der Daten in Frage. Aufgrund der häufigen Fehler bei den beiden Parsing-Analysen kann es zu Fehlinterpretationen der gelieferten Daten durch die Thematizer-Benutzenden kommen.

Trotz der genannten Schwierigkeiten sind die Ergebnisse hinsichtlich der Identifikation und Klassifikation der markierten Thematypen aufgrund der hohen F_1-Werte vielversprechend. Die Klassifikation der markierten Thematypen in mehrere semantischen Unterklassen stellt ebenfalls eine Bereicherung der thematischen Theorie dar. Diese Kategorisierung wurde zwar in früheren Forschungsarbeiten zum Teil bereits vorgenommen, jedoch fehlte bislang die automatische Bestimmung der markierten Thematypen und ihrer semantischen Unterklasse in computergestützten Thema-Rhema-Analysen. Da Thematizer es ermöglicht, markierte Themata mit hoher Genauigkeit zu identifizieren und ihren semantischen Beitrag zur kommunikativen Aussage zu bestimmen, können Schreibende durch thematische Analysen tiefere Einblicke in die semantische und logische Entwicklung ihres Textes gewinnen.

Neben dieser Erweiterung der thematischen Theorie stellt die Einbeziehung unmarkierter Themata bei Vorhandensein von markierten Themata eine weitere Abweichung von den bisherigen theoretischen Ansätzen dar. Die Beibehaltung der gleichzeitigen Realisierung von markierten und unmarkierten Themata untermauert deren Informationsgehalt als gegebene Diskursthemen, die nicht zur Weiterentwicklung des Diskurses beitragen, sondern als etablierte Basis der kommunikativen Aussage fungieren. Damit geht auch einher, dass keine Thema-bezogene Information verlorengeht oder der falschen Diskursfunktion von Satzgliedern zugeschrieben wird.
Auch die intertextuellen Analysen von Thematizer stellen einen wesentlichen Fortschritt im Vergleich zu bisherigen Softwaretools und Forschungsarbeiten dar. Da Thematizer mehrere Texte gleichzeitig analysieren und Häufigkeitsdaten für die darin markierten Themata und Thema-Rhema-Gliederung ermitteln kann, können Erkenntnisse über die thematische Struktur, Entwicklung und Verteilung zwischen den einzelnen Texten gewonnen werden. Gerade diese

Funktionalität der Software kann für textlinguistische Untersuchungen zu Intertextualität und Textsorte von besonderem Nutzen sein.

Nicht zuletzt ermöglicht die Visualisierung der thematischen Analysen einen konkreten Einblick in die thematische Struktur der vom Nutzenden hochgeladenen Texte. Durch farbliche Hervorhebungen, die direkt in den Text eingebettet und angezeigt werden, sowie durch aggregierte Häufigkeitsdaten in den jeweiligen Abbildungen können sich Schreibende in kurzer Zeit einen Überblick über die Verteilung der vorhandenen Themata, Thematypen und Rhemata verschaffen. Solche Ergebnisse können ein vertieftes Verständnis der syntaktischen Varietät und des lexikalischen Ausdrucks im eigenen Text fördern. Außdem verdeutlichen sie die thematische Struktur und die diskursive Entwicklung des Textes. In einer solchen Visualisierung der Ergebnisse liegt eine potenzielle Zugänglichkeit für Schreibende zur thematischen Theorie und ihrer Realisierung im Text.

Die Ergebnisse der vorliegenden Dissertation und die daraus gezogenen Schlussfolgerungen ebnen somit den Weg für weitere textlinguistische Forschung über die thematische Theorie. Zum einen kann Thematizer dazu verwendet werden, um thematisch annotierte Korpora zu erstellen, die bisher fehlen oder nicht öffentlich zugänglich sind. Solche Korpora könnten weitere textlinguistische Analysen erleichtern oder zumindest als Grundlage für (inter-)textuelle Forschung zu thematischer Struktur, Diskursanalyse und Informationsgehalt in geschriebener Sprache dienen. Darüber hinaus können thematisch annotierte Korpora als Datensätze für die Entwicklung sogenannter Transformer-basierter Pipelines (vgl. Wolf et al. 2020) genutzt werden. Diese basieren auf der Methodik und den Prinzipien des maschinellen Lernens und benötigen aus funktionalen und methodischen Gründen tausende Datensätzen, um die thematische Struktur im Text präzise erfassen, modellieren und akkurat analysieren zu können.

Eine weitere Forschungsmöglichkeit bietet die Zusammenführung von thematischen Analysen mit rhetorischen Strukturen. Da sogenannte rhetorische Aktivitäten (vgl. Cloran 1995: 362; 364-365) anhand des propositionalen Gehalts kommunikativer Aussagen generalisiert werden können, bereicherte die Zusammenführung von thematischer Struktur (abgeleitet aus Thema-Rhema-Gliederungen) und rhetorischer Struktur (abgeleitet aus dem propositionalen Gehalt der Aussagen) als Informationssysteme die Entwicklung und Analyse des Diskurses im Text.

Zukünftige Forschungsvorhaben sollten schließlich das oft vernachlässigte Rhema näher untersuchen. Da sich die meisten Forschungsarbeiten in erster Linie dem Thema und seiner inhaltlichen und diskursiven Entwicklung widmen, bleibt der Diskursbeitrag des Rhemas bislang eher zweitrangig. Die Kontextualisierung des Diskursbeitrags eines Rhemas fast ausschließlich in Bezug auf nachfolgende Themata unterstreicht die eher sekundäre Berücksichtigung des Rhemas in der Thema-Rhema-Gliederung. Das Thema fungiert als strukturelle Einheit gegebener Diskursthemen, als semantisches Feld für Agens, Prozesse und Sachverhaltsbestimmungen und als Rahmen des zu entwickelnden Diskurses. Im Gegensatz dazu werden die diskursiven und strukturellen Funktionen von Rhemata lediglich als Entwicklung der kommunikativen Äußerung durch neue Diskursthemen erwähnt. Eine genauere Untersuchung der Entwicklung des propositionalen Gehalts von Rhemata könnte, ähnlich wie bei der Thema-Rhema-Struktur, weitere Theorien zu den strukturellen, semantischen und diskursiven Eigenschaften des Rhemas liefern.

ACKNOWLEDGEMENTS

This dissertation would not have become a reality had it not been for the continued support, motivation, excitement and passion that I received from my advisor, Dr. Sanchez-Stockhammer. I cherish the work done together throughout the unfolding of this project, her invaluable insights, and not least the personal and professional input she offered at every stage of my doctoral work.

I am also very grateful for the assistance from my second advisor, Dr. Kunter, who provided key insights into the computational aspects of the dissertational work.

I would finally like to extend my sincerest gratitude to friends and family: To Pete, Katharina and Antonia, who continued to motivate and push me through the thick and thin of the project; to Claudia, Christian and Andrea, who aided in the editing and drafting process of the dissertation; and to my family, who stood by me in all aspects of my doctoral work over the years.

Table of Contents

List of Figures

List of Tables

Figures and Tables in Appendix A

Chapter 1 – Introduction

How language users construct sentences may appear to be a trivial task. After all, it simply requires appending one word after another to create the intended meaning. Yet, this simplicity belies the multitude of factors that underlie even the most basic of sentence constructions. Is the sentence being uttered in a spoken or a written setting? Which social factors, e.g., gender, age, socioeconomic standing or educational background, apply to the language user? Does the user adhere to or consciously deviate from language norms and conventions in their speech? What cognitive processes occur before and during language production that may affect the ultimate outcome of the utterance?

One question in particular that often haunts writers presented with a blank page or speakers in an ice-breaker situation is where to start. Knowing what to say is only part of the equation; being able to put thoughts to paper or orally is another hurdle unto itself. Consider the first sentence of this dissertation *How language users construct sentences may appear to be a trivial task.* This is by no means the only way the sentence could have been constructed, as shown in the following:

(1) The way in which language users construct sentences may appear to be a trivial task.
(2) * It is the way in which sentences are constructed that may appear to be a trivial task.
(3) * Trivial is what constructing sentences may appear to be.
(4) The construction of sentences (by language users) may appear to be a trivial task.
(5) ? There may appear to be a triviality behind how language users construct sentences.

These reformulations largely employ the same lexis as the original; however, either the order has been changed or additional constructions have been added such as the reformulation of the question word *how* as a relative clause together with *the way* in (1), the cleft in (2), the fronted element *trivial* in (3), nominalization and passivation of the subject in (4), or the existential *there* in (5). Considering each of these reformulations, they express the same propositional content: constructing sentences seems trivial. Pragmatically, however, i.e., the contextually implied or intended meaning, they differ.

The context of the original sentence and its reformulations is a formal, academic and empirical text in the form of a dissertation. Secondly, the context is further delimited by the fact that the sentence is the first of the entire text. Therefore, the sentence functionally sets the stage for the discourse to unfold. Functioning as the first sentence of the text therefore precludes lexical and textual reference to anything before its instantiation. In other words, the lexis that makes up the sentence does not point back to a previous sentence, only to abstract concepts initially outside of the text (here, constructing sentences and its supposed triviality).

Realizing the first sentence as a rhetorically emphatic cleft or fronted structure thus renders reformulations (2) and (3) infelicitous, as indicated by the asterisk at the beginning of their sentences. Further, sentence (5) would be questionable in its use, hence the question mark before the reformulation. What substantiates the contextually incorrect use of (2), (3) and possibly (5) is the function that the syntactic structures in these sentences afford. The cleft structure (e.g., *it is important that…* or *it is the argument outlined here that…)* in (2) and the fronted adjective in (3) have a corrective or emphatic function. Their employment assumes propositional information that has been explicitly realized (lexically, semantically or pragmatically) beforehand. If nothing comes before sentences (2) or (3), then their emphatic or corrective function becomes unjustified. Similarly, existentials as in (5) are commonly used to introduce new topics, but their introduction typically stems from topics previously realized in

the text (Halliday & Matthiessen 2014: 308). Again, if the reformulation in (5) were the first sentence, there would be no previously established topic upon which it could be based. Since neither sentence (1) nor (4) conflict with such pragmatic conditions, their use as potential reformations of the original first sentence is justified.

This brief exercise in sentence construction and reformulation refutes the assumption that sentence construction is anything but simple. The underlying grammatical principles of a given language certainly govern where sentence constituents must or should appear. However, lexicogrammatical requirements are by no means the only determining factor. If that were the case, then the grammatically correct constructions of (2), (3) and (5) would not be problematic.

A factor closely linked with the lexicogrammatical conditions of a sentence is that of word order. Particularly in English, word order is crucial as it determines the grammatical case of the words to be realized. This stands in stark contrast to languages such as German or French, which have case markings in addition to conventionalized word order. Further, the position of sentence constituents extends beyond determining grammatical case in English. Sentence constituents increase in informational weight towards the latter half of a sentence in standard declarative sentences in English, although exceptions abound (Quirk et al. 1985: 1391-1392). This means that discourse topics to be understood as prominent or of particular relevance tend to appear in the latter half of a sentence.

More specifically, the sentence constituents before the finite verb, typically reserved for the grammatical subject and fronted adverbials or adjuncts in standard, declarative sentences in English, contribute least to discourse development. In (1) and (4) above, these equate to *the way in which language users construct sentences* and *the construction of sentences*, respectively. On account of their lesser informational weight, they establish the foundation of the discourse message. Conversely, the elements that constitute the predicate (*may appear to be a trivial task* in both (1) and (4)) reflect the greatest informational weight within the sentences. In fact, it is the finite verb that traditionally functions as the border between less relevant or previously established information to the left of the verb and more relevant, discourse-developing information to the right (Firbas 1992, Adam 2013).

This bifurcated approach to modeling language at the sentence level forms the theoretical framework behind thematic theory and is the focal point of the present research. The theoretical underpinnings behind thematic theory were originally conceptualized by Weil in the late 19th century to express how word order is intrinsically linked to discourse development and the resulting communicative message (Weil 1978; originally published in French in 1844). As outlined in Chapter 2.1, Weil claims that language users orient the core of their message around topics that have already been mentioned in discourse. These topics establish the familiar or so-called 'known' information, upon which the user can expound. In the literature, familiar discourse topics are afforded the information status of GIVEN upon being explicitly realized within the discourse. As these topics do not drive the discourse further but rather form the foundation of further exposition in discourse, they carry comparatively little informational weight. Instead, they simply serve as a repeated starting point while progressing from one statement to the next. Such repeated and discursively established topics are known as the **theme** and are realized sentence initially.

In contrast, topics that are introduced for the first time in the discourse and realized within the predicate of a sentence constitute the so-called **rheme**. Since the propositional content that appears in the rheme is presented to the recipient as newsworthy or notable, it carries the most

informational weight and relevance in the discourse message.[1] Accordingly, the information status of the rheme is generally NEW information. Discursively, the rheme is responsible for moving communication forward through expansion, contradiction, contrast, comparison or any other logical means with respect to the theme. It answers the questions, "How would I like to specify the discourse topic presented in the theme of this statement?" or "What information would I like to add to the discourse topic presented in the theme?" In doing so, the speaker takes the limitless number of discourse topics they could use to talk about the theme and formulates a specific statement about it. Through the theme and rheme, both the propositional content and the information status of a sentence can be split into the thematic GIVEN and rhematic NEW.

Just as the finite verb functions as a pivot for the distribution of the sentence constituents' informational weight, it also functions as the boundary between the theme and the rheme in a sentence. Considering reformulation (4) again, restated in Figure 1-1, the grammatical subject *the construction of sentences (by language users)* constitutes the theme. The remaining sentence constituents, *may appear to be a trivial task,* then form the rheme.

THEME	**RHEME**
The construction of sentences (by language users)	*may appear to be a trivial task.*

Figure 1-1: Separation of sentence constituents into theme and rheme by using the finite verb, here the modal may, *as the boundary marker between the two.*

Through sentence-initial realization as the grammatical subject, the theme *the construction of sentences* establishes the starting point of the message, i.e., the foundation of the statement that the author wishes to make. It is not strictly necessary for themes to be both sentence initial and the grammatical subject; however, they must fulfill one of the two conditions. In the example sentence from Figure 1-1 then, of all the statements to be made about *the construction of sentences*, it is the presumed triviality of that task that the author then decides to discuss. Through realization of the rheme with this specific propositional content, the author moves the discourse forward on the basis of the theme. Even if the rheme were formulated differently, such as *varies from language to language*, it would maintain its rhematic status regardless; only the propositional content would change. In the same vein, the function of the rheme, namely that of pushing forward discourse, would also remain the same.

Over time, the treatment of theme and rheme as encoded by word order, informational weight and discursive function split into two schools of thought: functional sentence perspective and systemic functional grammar. Functional sentence perspective was represented by the Prague School of Linguistics, to which Weil, Mathesius, Firbas and Daneš belonged. Representatives of the systemic functional grammar school of thought are Halliday, Firth and Matthiessen. Both schools of thought share the term *functional*, which stresses the role that context plays in language use. Just as contextual cues were considered in ruling out problematic formulations in (1) – (5) above, realizing language as a conscious choice with motivated reasoning underlie both functional approaches to language (Davidse 1987; Christie & Martin 2010: 5; Derewianka 2011: 2). The choice that a language user has in realizing their own statements is critical to both schools and suggests a complex systematicity that permeates the nearly infinitesimal ways in which language can be produced. Despite the shared functional approach to language as a system, both schools came to develop divergent thematic models and corresponding terminology.

[1] It must be stressed that this is not unilaterally the case. There are a number of exceptions, including clefts, fronted elements and, in this work, thematic elements that appear before the grammatical subject of the independent clause. These exceptional cases and this work's divergence from the traditional approach to treating fronted thematic elements are discussed in greater detail in Chapter 2.6 and Chapter 4.3.

The greatest difference to emerge between both schools is in their determination of the informational weight that the theme and rheme contribute to the overall discourse. Within the functional sentence perspective framework, GIVEN and NEW elements are determined with the so-called communicative dynamism model, which was first put forward by Mathesius but formalized in Firbas' work (Adam 2013: 39). Communicative dynamism employs scales that reflect the syntactic realization of a statement in order to allocate varying degrees of informational weight to each sentence constituent; in doing so, the degree to which the sentence constituent contributes to the discourse message and its development can be traced (Firbas 1964b: 270). Elements which contribute less to the discourse are afforded thematic status and are most commonly conflated with the grammatical subject and sentence-initial constituents. The lower informational weight then affords the theme GIVEN information status. On the other end of the spectrum, elements with greater informational weight according to communicative dynamism are defined as the rheme and achieve NEW information status.

The systemic functional grammar approach similarly attributes informational weight to the syntactic realization of sentence constituents. Sentence elements in front of the finite verb are thematic, as already mentioned, and typically have GIVEN status. Contrary to the functional sentence perspective, however, when exceptional syntactic structures are employed, even thematic elements may be denoted as NEW information. For example, in reformulation (3) from above and restated as (6) in the following, the adjective *trivial* is fronted.

(6) Trivial is what constructing sentences may appear to be.

By breaking from the conventional SVO word order of English, the thematic *trivial* achieves NEW information status whereas the remaining rhematic constituents become GIVEN. As such, the systemic functional grammar perspective on thematic theory does not always conflate the theme with GIVEN and the rheme with NEW. Instead, information status shifts between both the theme and rheme depending on how they are realized syntactically and whether they follow standard word order and realizational patterns. While communicative dynamism can account for structures such as (6) through its use of various scales, the systemic functional grammar approach simplifies thematic analysis by relying on syntactic realization patterns alone to determine the theme and rheme. Despite this difference in approach, both schools have reciprocally contributed to the furthering of the thematic paradigm, and contemporary researchers typically take inspiration from both schools for their theoretical approach to thematic research.

Where the Prague School of Linguistics pushed forward thematic theory perhaps the most was with Daneš's (1974) models of so-called thematic progression, which was adapted and expounded upon by adherents of systemic functional grammar. So far, themes and rhemes have been considered at the sentence level but not across sentences clusters. If text is to be seen as a stringing together of phrases, clauses and/or sentences, it can be assumed that these text passages build upon one another in some fashion. Considering the text of this dissertation, the sentences were not haphazardly thrown together in an attempt to posit sentence structure with respect to thematic theory. Instead, new discourse topics in the text were presented against the backdrop of previously established ones; elaborations and examples were given to provide further detail; and overt sign-posting devices such as logical connectors were employed to facilitate the flow of information across sentences. This structuring of the text is thus reflected through the deliberate selection of themes (and rhemes) as the discourse unfolds, which is embodied in the concept of thematic progression.

Generally speaking, a text that has at least more than one clause – dependent or independent – has thematic progression. Consider the following newspaper excerpt taken from the Detroit News, whose themes are in bold:

> **Hurricane Ian** turned streets into rivers and blew down trees as it slammed into southwest Florida on Wednesday with 150 mph (241 kph) winds, pushing a wall of storm surge. **Ian's strength at landfall** was Category 4 and tied it for the fifth-strongest hurricane, when measured by wind speed, to ever strike the U.S. (Gomez-Licon 2022)

Hurricane Ian, which is the main topic of this news article, is used sentence initially and as the grammatical subject of the first sentence. As such, it forms the foundation of the statement and assumes that the reader is aware of this topic or at least of what a hurricane is. The effect that Hurricane Ian had is outlined in the rheme that follows to develop the discourse through the inclusion of NEW discourse topics. Next, the author chooses to continue the discourse around *Ian* by starting the second sentence with *Ian's strength at landfall*. The repetition of the same theme *Ian* across two sentences exemplifies the return to GIVEN information, i.e., information that has already been explicitly mentioned or realized in a previous sentence. When progressing from the first sentence to the next, the author could have instantiated anything from the first sentence, either from the theme or the rheme, as the theme of the second sentence. For example, instead of using *Ian* as the same theme in the second sentence, the author could have further elaborated on the US state of Florida, the ramifications of the trees being blown down or even the speed of the winds. The important note here is that either GIVEN (thematic) or NEW (rhematic) information from one sentence is then realized as GIVEN (thematic) in the subsequent sentence. Again, the reason for this is that the GIVEN information, the theme, represents the foundation or starting point of the message that has already been explicitly mentioned previously and thereby familiar to the reader.

Instances whereby the same theme is realized from one sentence to the next is known as constant continuous progression, as originally postulated by Daneš (1974: 118). Had the author re-instantiated the rheme from the first sentence as the theme in the second sentence, then simple linear progression would have been present. These two are by no means the only ways in which thematic progression ensues, as Chapter 3.1.1 outlines. Regardless of how themes and rhemes are developed across sentences, it begs the question of the importance of thematic progression in the first place. What meaning does the author's decision behind instantiating the same theme, but not the rheme, hold? How does this choice of theme or rheme inform the unfolding of discourse in text?

Firstly, the matter of choice returns as an important cornerstone of a functional and systemic approach to language. Both adherents of the functional sentence perspective and systemic functional grammar subscribe to the notion that a contextually motivated choice, whether conscious or not, creates meaning (Repka 2021: 167). Depending on the discourse-specific goals in a given context, there are certain linguistic choices that will contribute to the fulfillment of said goals. As Derewianka states, "[a]t the level of specific situations […], the [language] model indicates how choices from the language system are influenced by certain features of the situation" (2012: 132).

In the case of the newspaper article above, it serves to fulfill the purpose of informing a wider audience of a recent natural disaster. Therefore, more straightforward thematic progression patterns are employed, i.e., continuous progression, which simplify the reception of the information being transmitted through explicit repetition of the same sentence constituents. Had the rheme from the first sentence been realized as the theme in the second, that would have

resulted in a repeated shift in topics. Over longer pieces of text, that demands greater cognitive involvement on the part of the reader since they must retain a wider range of information and newly introduced discourse topics. Instead, maintaining the same topic through continuous re-instantiation as the theme (*Hurricane Ian*) eases comprehensibility and facilitates greater accessibility to the information. This invariably fulfills the informative goal that newspaper articles have more readily. In short, the discourse context informs the choices languages users make in readily achieving discourse goals. Here, one choice the language user is presented with is the development and presentation of information across sentences and throughout a text. Thematic progression patterns thereby reveal the speaker's linguistic choices in the process of meaning making.

The second reason for the importance of thematic progression patterns is that they can reveal characteristics of a text type. Similar to register, lexical density and syntactic complexity, thematic progression patterns as an intertextual characteristic help to shape the conventions of text types (North 2005: 432). These conventions are reflected in text genres, which embody the respective expectations, i.e., discourse goals, of a given text type. As Figueiredo points out, it is essential for language users to follow certain (but not necessarily all) conventions of a text type; this allows the text recipient to confirm the degree to which a text's discourse goals have been met and, in turn, categorize the text as a specific text type (Figueiredo 2010: 130). For example, a greeting and a closing, frequent use of first- and second-person pronouns and a lower lexical density are conventional intertextual characteristics of a personal letter (cf. Connor & Upton 2003). If technical jargon, overuse of nominalization and more subordination than coordination were employed instead, the reader may have greater difficulty in qualifying the text as a personal letter and may receive the text with divergent expectations. As such, active and deliberate employment of these characteristics in text production aid the language user in identifying the text as such.

In the same vein, text types and genres may reflect thematic progression patterns that can be used to help identify them accordingly. As shown in the brief example of the newspaper article above, continuous thematic progression was evident for reasons of ease of comprehensibility and information accessibility (Francis 1989: 212). Fries reported similar findings in his examination of narratives, albeit split between young adult and adult readers: young adult narratives appeared to generally reflect a greater occurrence of continuous thematic progression, whereas adult narratives showed more frequent simple linear progression (1995: 353). Similar findings on the relationship between text type and thematic progression pattern were reported by Martin (1993), Swales (1990), Downing (2001) and Berry (1995).

That being said, this stance is not shared amongst all linguists: Loftipour-Saedi & Rezai-Tajani (1996) and Mauranen (1993), for example, recognize thematic progression patterns as a characteristic of texts, but one that neither defines nor unequivocally qualifies a text as belonging to a certain text type. Instead, they merely fulfill a secondary texture characteristic in specific sections of a text alone. For many, it is less so the thematic progression *patterns* and rather the *types* of themes employed that contribute to a text type's characterization. Circumstantial themes, e.g., adverbial phrases, have shown greater employment in history textbooks as opposed to scientific texts (North 2005). Thematized places are employed frequently in guides, whereas thematized agents and times appear often in bibliographies (Enkvist 1987, Lavid 2000). Finally, thematized relational processes have been shown to be a characterizing aspect of editorials and letters (Francis 1989; see also Gosden 1992/1994, Rosa 2013, Jalilifar 2009). Regardless of whether the theme type or thematic progression pattern contributes more readily to a text type's characterization, both camps argue that their function

is a rhetorical one. In other words, theme types and progression patterns ultimately serve the fulfillment of a text's discourse goals of a given context, i.e., text type.

Returning to thematic progression patterns specifically, one final advantage can be found in the writing process itself, particularly for non-natives. If the thematic progression patterns of a text can be assumed to indicate how the language user develops discourse topics from one sentence to the next, these patterns will ultimately reflect the text's structural and logical development. A trace is then left from sentence to sentence, whereby the theme in one sentence is derived from a previous theme or rheme so long as there are no thematic breaks, i.e., the lack of thematic progression between sentences. The presence or absence of thematic progression between sentences, i.e., thematic traces, can serve as a tool or metric during the writing process.

By consciously instantiating subsequent themes on the basis of previous themes or rhemes, writers can more readily ensure cohesion across sentences. Cohesion can more effectively be achieved since themes can be realized through coreferential devices, such as synonymy, parallelism, lexical entailment or paraphrase. While thematic progression patterns may not guarantee coherence, they may reinforce either the logical development of a text when maintained or potential gaps in logic when broken (Rose 2001: 3-4; Jingxia & Li 2013: 120-121). For example, if an author were to write the following sentence with the theme in bold:

(7) **Little evidence** has been provided for the arguments made in the lawsuit.

they could follow it up with any of the following to maintain the logical development of their text by means of constant continuous progression for (8) and (9) or simple linear progression for (10) and (11):

(8) **Such evidence** would be required to justify their claim.
(9) **If provided**, then the arguments would have greater merit.
(10) **These** therefore have no grounds in the present case.
(11) Since **these kinds of legal disputes** are thoroughly scrutinized, it would be of little surprise that the law court would require substantial evidence.

The bold portions of the sentences indicate the theme that has been instantiated on the basis of the theme or rheme in the previous sentence. Note that *evidence* has been elided in sentence (9), which would thereby account for the cohesion reinforced through the continuous thematic pattern between the two sentences. As each of the themes is realized using cohesive devices (lexical repetition in (8), ellipsis in (9), coreference through the demonstrative pronoun in (10), and paraphrase in (11)), the structural integrity is maintained across the sentences. Further, the re-instantiation of a previous theme or rheme coreferentially accounts for the maintenance of the coherence, and thereby logic, between both sentences.

These four sentences could then be contrasted with the following two which break the coherence between the two sentences (the original sentence has been provided in both again for clarity):

(12) *** Little evidence** has been provided for the arguments made in the lawsuit. **A helpful tool** is the argumentation.
(13) *** Little evidence** has been provided for the arguments made in the lawsuit. **Contrarily, significant effort** is required to address companies' quarterly reviews.

Again, the asterisk indicates a problematic formulation, here for reasons of lacking coherence. In (12), the gap in logic emerges due to NEW information being introduced as the theme in the second sentence. In this example, the topic of *a helpful tool* can be qualified as NEW due to the indefinite article *a* (McCabe 1999: 172) and because it cannot be explained by means of a cohesive device. Further, it has not been explicitly introduced as a discourse topic in previous sentences. Finally, the NEW discourse topic of *arguments* in (12) is realized as the GIVEN discourse topic *argumentation* in the rheme of the second sentence through the definite article *the*. Hence, the NEW information is realized in the incorrect place (the theme) of the sentence, just as the potential coreferencing and now GIVEN constituent *argumentation* is incorrectly realized in the rheme. An incorrect reversal of theme-GIVEN followed by rheme-NEW is particularly common in non-native English writers, although it is not entirely unique to that group alone (Ahmed et al. 2015). As for (13), while the cohesive device *contrarily* is used to establish cohesion between the two sentences, coherence fails due to entirely NEW topics being realized both thematically and rhematically. While NEW discourse topics are appropriate within the rheme, the theme must include at least one instance of GIVEN information in standard SVO English sentences to achieve successful thematic progression.

These examples briefly highlight that neither grammatically correct sentences nor established cohesion can account for coherence across sentence boundaries. However, with the help of thematic progression, writers can identify potential gaps in the structural and logical development of their text. It further highlights the role that GIVEN vs. NEW information plays in sentence construction, particularly since this differs from language to language and, in the case of English, is highly dependent on word order (Firbas 1964a: 112; Arnold et al. 2013: 404).

These three key facets of thematic progression patterns – how they reveal the text's method of development through thematic selection, their ability to indicate text type characteristics, and their use as a writing tool for identifying potential gaps in the text's structure or logic – can be a decisive asset for language users of any level or background. However, being able to trace the thematic progression that a text has can prove difficult, even for those familiar with the concept. This issue is further complicated by the numerous exceptional cases that thematic theory possesses.

To date, there has been a wealth of research on teaching thematic theory for tracing thematic progression in one's own writing for both L1 and L2 writers. Both Downing (2001) and Downing & Locke (2006) have shown how Daneš's thematic progression patterns can be effectively used together with the discursive functions of the theme and rheme to structure text and achieve discourse goals. Whereas the theme fulfills a scaffolding role through continuous re-instantiation of GIVEN discourse topics, the communicative relevance of rheme with respect to the theme is its ability to orient the discourse around context-specific goals characteristic of a text type (Downing 2001: 6). North (2005) focused her attention on the relationship between thematic progression patterns and academic disciplines in undergraduate university papers. Findings indicate that arts students, compared to science students, more frequently made use of interpersonal themes on the basis of previous established argumentation to contextualize and frame their thematic exposition instead of presenting information simply as fact (2005: 449). Further, Moore (2006) and Hawes (2015) advocate leveraging GIVEN and NEW information in conjunction with thematic progression. The reason behind this is to make information presented in the text more accessible in terms of established topicality and direction of text development and structure (Moore 2006: 11; Hawes 2015: 98). Finally, in translation and teaching translation, Jalilifar (2009), Williams (2009) and McCabe (1999) stress the importance of exploiting theme/rheme and GIVEN/NEW information from a text in one language as an aid in translating the text into another language in accordance with that language's GIVEN/NEW pattering. For

example, while GIVEN most commonly appears first in English due to word order restrictions, this does not apply to languages with more flexible word order, such as German or Russian. Being aware of these systemic differences between languages in terms of theme/rheme and GIVEN/NEW structures can thus aid in the transferal of information across languages while maintaining idiomaticity in the target language (Jalilifar 2009: 107).

Despite the theoretical advancements made in the thematic paradigm for use as a writing tool, there remains a poignant drawback to thematic analysis: it is time consuming. Thematic analysis requires the dissection of every sentence, which, even for shorter compositions, can become a major undertaking. This first involves the identification and subsequent confirmation of the themes and rhemes of each sentence. Afterwards, this information is used to determine the thematic progression patterns that may be present (or markedly missing) throughout the text. Finally, as will be made apparent in Chapter 2.5, not all themes are made the same: they can be categorized into textual, interpersonal or topical, according to Halliday, and each serve a different textual, logical or structural purpose. The more fine-grained an analysis the language user desires, the more time it will require.

For that reason, automating the analysis by computational means has become a recent step in bringing thematic analyses to the 21st century. By and large, most recent research on thematic analysis has been for the purposes of text segmentation (Popping 2000, Hotho et al. 2003, van Atteveldt et al. 2021), theme identification using machine learning algorithms (Lavid 2000, Moens 2007, Hajičová & Mírovský 2018, Xi et al. 2020) and text classification and information retrieval (Steinberger & Bennett 1994, Kappagoda 2009). Most commonly, however, these approaches centered solely around the automated identification of discourse topics in text, excluding an analysis of individual themes, rhemes and thematic progression.

Three separate groups of researchers who examined either thematic theory or thematic progression from an automated and computational perspective are Schwarz et al. (2008), Park & Lu (2015) and Domínguez et al. (2020). Schwarz et al. was one of the first groups to conceptualize an automated approach to identifying themes by using the Stanford PCFG parser and rule-based thematic patterns as originally defined by Halliday (2004: 65-81). With their Hallidayan approach, they focused on simple themes alone, meaning that only singular theme types were considered and extracted from the analyses. Park and Lu (2015) expounded on this work by identifying multiple types of themes (textual, interpersonal and topical) as postulated by systemic functional grammar. Furthermore, their software, Theme Analyzer, was programmed to identify sentence mood, the theme/rheme boundary and the syntactic role of the theme, such as complement or adjunct. Finally, Domínguez et al. (2020) focused on thematic progression by means of theme identification using the Spacy parser and a rule-based approach. Their software, ThemePro, not only identified and visualized themes and thematic progression, but also syntactic trees and coreference chains. While they created and tested their work for the written mode, it was conceptualized for eventual use in spoken speech. Their theoretical and conceptual approach to thematic analysis deviated slightly from that of Halliday, which resulted in themes being considered one class.

All three bodies of research represent a significant contribution to the study of thematic theory via computational means. Particularly Domínguez et al. (2020) have made an automated analysis of thematic progression much more accessible thanks to providing their source code on Github. Otherwise, to the best of the author's knowledge, there are currently no publicly available tools or theme-tagged corpora that researchers or linguists can use for an automated analysis of their texts. Furthermore, limitations to ThemePro and other automated tools are their deficient visualization of the analytical data, their inability to analyze multiple texts for

intertextual analyses, their inability to download the resulting text analyses for tagging or general corpus use, general statistical data, such as the number of themes and thematic progression patterns, and a differentiation between theme types.

Against the backdrop of previous theoretical and computation approaches to thematic theory, the current work aims to achieve two research goals: firstly, to identify and overcome deficiencies in previous thematic analysis software and in contemporary models of thematic theory; secondly, to deliver a web-based feedback tool, called Thematizer, that makes thematic theory accessible to writers through its operationalization by computational means. Here, operationalization means the application of the thematic framework as the theoretical underpinning behind Thematizer. Accessibility to thematic theory is then understood as a function of operationalization: The more accurate Thematizer's thematic analyses are, the more successful its operationalization of thematic theory. This, in turn, increases the user's accessibility to the thematic paradigm.

Starting with the deficiencies identified from previous thematic theory models, the first is the exclusion of unmarked themes in the presence of marked themes and the further classification of marked themes into their functional categories. As thematic models that follow the Hallidayan approach relegate unmarked topical themes as grammatical subjects to the rheme if a marked topical theme is realized, GIVEN discourse topics become obfuscated in corresponding thematic progression analyses (cf. Figure 1-2). Maintaining both a marked and unmarked theme in thematic constituent analyses, as done in the present work, ensures that GIVEN discourse topics are not subsumed under the NEW rheme; instead, they retain their thematic status.

Hallidayan Approach	MARKED THEME				RHEME		
Present Work	MARKED CIRCUMSTANTIAL THEME				UNMARKED THEME		RHEME
Text	*In*	*the*	*general*	*report*	*several*	*participants*	*were identified .*

Figure 1-2: The present work's deviation from the traditional Hallidayan approach to marked and unmarked themes, such that unmarked themes are included in analysis even when a marked theme is realized. This ensures that previously established, i.e., GIVEN, discourse topics retain their thematic status.

The second core deficiency identified is the further delineation of marked theme types and subclassifications. While previous research has considered marked themes in terms of theme roles or metafunctions, such as topical or interpersonal, the limited marked theme classifications fail to reflect the syntactic diversity and logical functions that marked themes have. In other words, marked topical themes can range syntactically from complex hypotactic clauses to varied prepositional phrases and even to simplex temporal noun phrases such as *yesterday*. As such, further analysis of marked themes' syntactic and functional categories is required to reveal their contribution to the structural or logical development of the text. The present work thereby classifies marked themes into additional syntactic categories, which are then broken down even further into their semantic subclass, e.g., TEMPORALITY for the circumstantial theme *in the past* or CONTINGENCY for the hypotactic theme *if necessary*. Additional marked theme types and their categorization into their semantic subclasses affords richer detail to Thematizer's thematic analyses without manual analysis on the user's part. For writers, this means that they are provided with information on the syntactic diversity (or heterogeneity) of their writing; for researchers, they can gain insight into overall thematic progression in a given text while simultaneously delving into the specific distribution of marked themes, their realization patterns and semantic subclasses.

Additionally, this work addresses a key drawback to thematic analyses previously mentioned: the effort and time required for analysis. In order to reveal intertextual characteristics of a text type at a statistical level, vast amounts of text are required for statistical representativeness. Analysis is thereby limited by the number of texts to be analyzed manually. Through automation, text analysis can be expediated and without formal training on thematic theory. While the quality of the analyses must be ensured, the ability to rapidly produce data on thematic analyses could be a boon to generating tagged corpora and pertinent texture parameters. This is necessitated even further in machine-learning environments since even greater amounts of tagged input are required to produce reliable and comprehensive language models. The software developed in this work can function not only as a facilitative tool to writers of any background but also as a computational steppingstone to advancing text analysis in the linguistic community and digital world.

Improvements to previous iterations of thematic analysis software formed another core impetus behind the present work. The visualization of thematic analyses was identified as particularly deficient as results were presented as abstract numerical values without the use of the user's original text. Therefore, how thematic progression or thematic constituents were realized and developed remained blurred in the analytical output. The limitation of one text per analysis in previous software iterations also prevented intertextual analyses, which are pivotal in establishing intertextual and texture characteristics of text types. The remaining deficiency identified was the user's lacking accessibility to the analytical data. While the software presented the results from the thematic analyses, users were not given the option to export the output for their own use. The present work therefore sought to overcome each of these key deficiencies through the development of Thematizer.

The development of Thematizer as an automated tool for thematic analysis is the product of the present dissertation and research on thematic theory. The theoretical, programmatic and functional framework that underpins Thematizer forms the core of the present research. This framework is addressed together with the key findings from a computational approach to thematic theory throughout the remainder of the dissertation.

Chapter 2 traces the origins and development of thematic theory, starting with a brief treatment of Weil and his conceptual contribution to thematic theory. Further expansionary work on thematic theory from the Prague School follows, culminating in Daneš's conceptualization of thematic progression. Next, the systemic functional grammar model of thematic theory is presented with a breakdown of thematic constituent analysis according to Halliday. In the penultimate section of Chapter 2, exception cases for how theme and rheme constituents are determined according to both schools of thought are addressed. Finally, Chapter 2 concludes with a summary of the conceptual definitions of theme, rheme and their identification in text from Prague School and Hallidayan perspectives.

Chapter 3 considers how themes can be used as a structural and methodological tool for written discourse. It specifically considers the superordinate role that the theme plays in the theme/rheme dichotomy, and how semantic and pragmatic functions are reflected in its realization. Chapter 3 continues with a closer dissection of Daneš's original conceptualization of thematic progression, albeit from a text development perspective. Expansions to Daneš's original models from recent contemporary research are provided with explanations regarding their addition to the overall thematic progression framework. Next, thematic progression as a characteristic of a text's textuality and potential indication of text type is addressed. This then ties into contemporary research on thematic analysis from a computational perspective, which forms the concluding discussion of Chapter 3.

Chapter 4 marks the shift from previous research on thematic theory to the work conducted for the present dissertation. There, the formal research questions that drove the theoretical and practical work behind Thematizer are first outlined. Next, the concepts of theme, rheme and thematic progression are formalized. This includes a conceptual definition of each and a breakdown of all thematic realizational patterns employed for thematic constituent and thematic progression analysis. Chapter 4.5 begins the methodological discussion on the development of Thematizer by outlining the materials used to develop, train and test the software. Its program specifications, such as programming language and libraries, are treated thereafter.

Chapters 4.7 to 4.9 outline the analytical functionality that Thematizer was programmed with. These sections detail how the theoretical framework of thematic structure was operationalized via computational means. Specifically, the three core processing steps that Thematizer performs in its thematic analyses are explained: the identification of thematic and rhematic constituents, marked theme identification and classification, and classification of thematic progression patterns. Chapter 4 concludes with a presentation and explanation of Thematizer's web interface and analytical output that is delivered to users upon completion of the text analysis.

Chapter 5 covers the performative results that Thematizer yielded in its training and data validation. This is expressed in terms of parsing accuracy as a function of correct thematic constituent analyses versus erroneous and overlooked identification of thematic patterns. Prominent and recurring parsing errors are summarized as key error classes in Thematizer's parses and overall functionality in order to provide initial answers to the underlying research questions behind the present work. In this chapter, the results are presented with tentative interpretation of the causes behind the emergence of key error classes.

Chapter 6 then delves into a comprehensive interpretation of the key error classes, their effect on Thematizer's parsing accuracy and the reasons for their emergence. These discussion points are treated individually with respect to Thematizer's three core parsing tasks in order to elucidate the degree of successful operationalization of thematic theory. The findings presented in this chapter are contextualized around findings from other contemporary and related research on thematic theory from a computational and theoretical perspective. Taken together, this collection of findings informs the final answers, and thereby conclusions, to the present work's underlying research questions.

Chapter 7 concludes the dissertation with an initial summary of the key findings presented in Chapters 5 and 6. Here, final conclusions drawn from Thematizer's performative results are reiterated with the key error classes that emerged in each of its three core parsing tasks. Limitations to the present work are then outlined. Avenues of research that these conclusions and findings enable as suggestions for future work on thematic theory constitute the final discussion of the dissertation.

References to all figures and tables employed in this work have been provided at the beginning of the dissertation for ease of reference. Further, an appendix has been added after Chapter 7, which summarizes certain key error classes for Thematizer's training and test data set as well as detailed accuracy scores achieved in the three core parsing tasks.

Chapter 2 – Thematic Theory from Weil to Halliday

This chapter presents the concepts of theme and rheme from their initial conceptualization by the Prague School of Linguistics to their further refinement within the Hallidayan framework. The treatment of theme and rheme as a theoretical model for text analysis through the lens of both schools of thought will highlight their conceptual differences in the characterization and definition of thematic theory. The additional purpose of this theoretical background is to provide a comprehensive presentation of thematic theory from a conceptual perspective, which informed the theoretical models employed by the present work. In doing so, the theoretical foundation for the discursive functions of theme, rheme and eventually thematic progression can be laid.

Chapters 2.1 to 2.4 concern themselves with the adherents of the Prague School who first conceptualized and forwarded thematic theory. Specifically, theoretical contributions from Weil, Mathesius, Firbas and Daneš as Prague School linguists are addressed in brief. Chapter 2.5 then shifts to the interpretation of thematic theory from a systemic functional grammar perspective as formalized by Halliday. A summary of both schools' understanding of thematic theory is presented in Chapter 2.6, which is followed by an excursus of the treatment of exception cases in thematic constituent analysis in Chapter 2.7.

2.1. Weil

The topic of theme and rheme reflects a rich historical development from its initial beginnings in the 19th century. At that time, Weil, who laid the groundwork for thematic theory, made first attempts to divide utterances into two disparate yet interdependent parts: *le point de depart* and *le but de discours* or *l'énonciation*. Each of these constituents were later formalized as theme and rheme by Mathesius (cf. Chapter 2.2), who was inspired by Weil's treatment of word order, information status and discourse topic realization in text.[2] According to Weil:

> There is […] a point of departure, an initial notion which is equally present to him who speaks and to him who hears, which forms, as it were, the ground upon which the two intelligences meet; and another part of discourse which forms the statement (l'énonciation), properly so called. This division is found in almost all we say. (Weil 1884/Engl. translation 1978: 29)

Weil argues fundamentally at the syntagmatic level, stating that "word order is the order of ideas: 'general ideas' are stated before 'special ideas', the given information precedes the new information" (de Jonge 2007: 228). Statements are dissected into a bipartition of "psychological [i.e., grammatical] subject and predicate", whose syntactic relations Weil subsumes under 'objective movement' (Weil 1978: 29). The objective movement of an utterance is therefore its actual syntagmatic realization, i.e., how a statement is realized syntactically. Parallel to objective movement, Weil proposed subjective natural movement, which encapsulates the information status of a text's discourse. Subjective natural movement is achieved through a statement's point of departure ('*le point de départ*') and the goal of the discourse ('*le but de discours*', '*l'énonciation*'). It is the latter that is responsible for developing discourse on the grounds provided by the point of departure. This, in turn, reflects the informational character of a statement: the point of departure corresponds to GIVEN information and the goal of the discourse to NEW information (cf. Figure 2-1).

[2] See also Ammann 1928, who similarly employed the concepts of *thema* and *rhema.*

	GIVEN & Thematic	NEW & Rhematic
Subjective Natural Movement	point of departure (*le point de depart*)	goal of the discourse (*le but de discours*)
Objective Movement of Statement 1	In truth it is not **gold and silver**	which make life comfortable.
Objective Movement of Statement 2	Having **these metals** only	would make people very miserable.

Figure 2-1: Illustration of Weil's subjective natural movement and objective movement as a parallel expression of syntactic realization and information status of discourse topics and as inspired by the example provided in Weil 1978: 39. The terms given in bold are the GIVEN themes that are developed between both sentences. The terms 'thematic' and 'rhematic' are employed as a parallel to terminology used in thematic theory, although Weil did not make use of these terms himself.

Propositional content denoted as GIVEN specifically refers to discourse topics that have been mentioned previously within the text. Conversely, NEW information is propositional content that has been realized for the first time in the discourse and cannot be traced back to previous textual content within the discourse. GIVEN discourse topics are thus re-instantiated throughout discourse to establish a foundation for the communicative message that unfolds in juxtaposition with the NEW discourse topics realized as the goal of the discourse. In other words, the NEW discourse topics move the message forward as contextualized around GIVEN, i.e., previously established, discourse topics.

Weil contrasts these two systems, **objective** and **subjective** as one system, GIVEN and NEW as another, by illustrating their function in ancient as opposed to modern languages. Languages such as Ancient Greek and Latin were both characterized by free word order, which caused the first lexical item to be encoded as the point of departure in a statement. For that reason, any number of sentence-initial syntagma could be encoded as the point of departure irrespective of syntactic function. Using Weil's terminology, the subjective natural movement motivated a speaker's realization of an utterance with less consideration of the statement's objective movement due to free word order. In other words, instead of syntax governing the order of realization for discourse topics, information status via GIVEN as the point of departure and NEW as the goal of the discourse determined discourse realization (Weil 1978: 36-37). Since syntagma could be freely placed, greater precedence was given to the information status and thereby realization and reception of the communicative message.

Conversely, modern languages, Weil contends, tend to conflate the point of departure with the grammatical subject, which may be realized at various parts of an utterance. Should the subject be realized sentence initially, the statement would reflect minimally animated syntax and natural word order (e.g., statement 2 in Figure 2-1). Statements whose subject is realized medially or finally would then reflect "unnatural" or atypical word order and "the most animated syntax" (Weil 1978: 37). Weil defined such animated syntax via atypical word order as 'pathetic' ('*l'ordre pathétique*'). This is the equivalent to marked characterization in contemporary terminology due to the grammatical subject not being realized sentence initially. In such cases, the sentence-initial lexical item(s) is realized as the goal of the discourse instead of the point of departure. The result is NEW information appearing before GIVEN. For example, *The answer I didn't know* reflects this phenomenon. Here, *the answer* is fronted to provide a "vehicle of emotion" and hence a stressed or emphatic function (Weil 1978: 12). If adverbials, such as prepositional phrases or conjunctive adverbials, appear before the grammatical subject, the statement is not considered pathetic but highly animated as in statement 1 from Figure 2-1. In such cases, the sentence-initial adverbial might function as a text's point of departure to contextualize the GIVEN grammatical subject and NEW predicate that follow.

This section's brief treatment of Weil's contribution to the informational organization of discourse topics vis-à-vis syntactic realization served to highlight the interplay between information status and word order: GIVEN discourse topics form the foundation of the discourse as the grammatical subject and precede NEW ideas realized in the predicate and as the goal of discourse. Deviations from standard word order and the standard presentation of ideas from general to specific reflect the speaker's emphatic and emotive rhetoric as a reinforcement of information status and unfolding discourse. Weil's postulation of objective and subjective natural movement in language expression thus came to inform the parallel systems of discourse realization and information status in later models of thematic theory. Finally, his interpretation of encoding the grammatical subject with the communicative message's point of departure became a formative feature in subsequent linguists' demarcation of theme and rheme.

2.2. Mathesius

Mathesius' development of the thematic paradigm embodies major shifts in linguistics at the time of his research. The field was readily moving away from a primarily diachronic approach to language analysis. Instead, a synchronic, descriptive approach to language less engrained in prescriptivism came to shape linguistic research. At the time of Mathesius' writing, de Saussure released his seminal work on *langue* and *parol* (de Saussure 1988). This marked a break from traditional approaches to language toward a contemporary, scientific approach to understanding linguistic phenomena. Simultaneously, a functional approach to language began to emerge, particularly from the Prague School, to which Mathesius belonged. As outlined in his work *New Currents and Tendencies in Linguistic Research*, Mathesius argued that "[…] modern linguistics more and more takes the meaning or function as its starting point and tries to find out by which means it is expressed. This is the point of view of the speaker or the writer who has to find linguistic forms for what he wishes to express" (Mathesius 1985: 57). Thus, context plays an unequivocal role in communication as it aids in providing the foundational basis and impetus for the discourse. Within this framework, context is understood as a fundamental and indispensable function of communication that all language expression can be attributed to. Context thus affects the resulting lexical selection and syntactic patterns of the discourse that unfolds.

This functional, i.e., contextual, aspect of language and language systems came to shape the theory of functional sentence perspective, which identifies the sentence as the cornerstone of textual expression. Mathesius later defined this as "a communicative utterance by which the speaker assumes an active attitude to some fact or a group of facts" (Mathesius 1983: 124). Therefore, the sentence as the most basic form of linguistic expression need not only be defined in terms of lexicogrammatical constituents. Rather, it is construed through a "point of view" and "active attitude to some fact" (Mathesius 1983: 124). This definition allows for a much more plastic interpretation of what constitutes a sentence and its interpretability.

Context, point of view and active attitude to some fact are all subsumed under the terms foundation (Cz. *základ*) and starting point (Cz. *východisko*[3]), which correspond to Mathesius' definition of theme (Cz. *tema*). Further, the core (Cz. *jádro*) or enunciation (*l'énonciation*) of the utterance finds its culmination in the remaining constituents of the text, i.e., the rheme.[4] These concepts are then intertwined in all forms of communication: the language user initiates a common ground (point of departure), realized by means of the foundation or theme and

[3] It should be noted that Mathesius later dropped this term in favor of simply using *základ* and *tema*, both of which he used interchangeably.

[4] Further nomenclature for theme/rheme, particularly within the Prague school vis-à-vis GIVEN vs. NEW, is topic/focus (see Gundel 1988).

articulated further by the predicating rheme. The foundation and core of a discourse are ultimately reflected in the word order of the text that is produced. Mathesius systematized this interplay by observing syntagmatic, contextual and discourse factors and examining the order in which subjects and predicates were realized in a sentence according to specific contexts. These initial examinations in the case of Czech laid the foundation of Mathesius' work on the thematic paradigm.

Mathesius' research represents a turning point in the development of theme and rheme as "the notion of theme was clearly articulated by Mathesius […] and has been developed by members of the Prague school since then" (Fries 1995: 317). Building off traditional definitions of sentence constituents, Mathesius makes first mention of the theme-rheme dichotomy when defining communicative utterances as thetic or predicative (Mathesius 1975: 87). In either of these cases, the theme is followed by the rheme to form standard word order and thereby unmarked thematic realization. The theme is therefore that which has already been made clear or familiar to the text recipient and serves as the foundation of the discourse. The foundation or starting point of discourse is primarily motivated by the context in which the discourse and participating speakers are embedded. However, the speaker may make certain assumptions about the speech partner's background knowledge when creating discourse. False assumptions about what information a speech partner can access invariably lead to miscommunication and misunderstandings since the speaker attempts to refer to information the recipient lacks. For example, in (1), miscommunication may occur if the text recipient is unaware of Mike's party or who Mike is.

(1) Mike's party is supposed to start at 11 tonight.

Communication can break down if the speaker fails to mention required information previously, if the context does not provide sufficient explanation, or if the recipient was not meant to be provided with the information (e.g., if the recipient was not meant to be invited to the party).

Assuming these factors have been accounted for, (1) illustrates standard (or "subjective" as per Mathesius) syntactic word order as well as the standard order for thematic and rhematic constituents: *Mike's party* qualifies as the theme of the sentence, *is supposed to start at 11 tonight* as the rheme. This, however, is not the only way this utterance can be expressed. Depending on contextual or cotextual cues, it may be realized with what Mathesius calls "objective word-order" (1975: 94). In utterances with objective word order, the sentence initial constituent is realized by an element that is not the grammatical subject. In the example above, this could take any of the following forms:

(2) Tonight, Mike's party is supposed to start at 11.
(3) At 11 tonight, Mike's party is supposed to start.
(4) Tonight at 11, Mike's party is supposed to start.

Mathesius attributes deviations to standard word order to "the novelty or lack of novelty of the notions expressed by the different sentence parts, emphasis and emotion, the content and complexity of the expression" (Mathesius 1983: 126). Thus, in (2) – (4), the explanation for the temporal adverbials realized as sentence initials could be found in their emphatic and/or emotive expression. The formulation in (2) could further be explained with Mathesius' reasoning that the two adverbials increase the complexity of the utterance and could be split for economy or processability reasons. Since *tonight* and *at 11* are to be understood as novel information, which would merit their placement before the grammatical subject, they would reflect rhematic status. Thus, in (2) – (4), Mathesius attributes thematic status to *Mike's party*

with all remaining constituents receiving rhematic status. This means that multiple rhemes can be present in a single sentence and split by the inserted grammatical subject theme. Although Mathesius warns against conflating the grammatical subject with theme and the predicate with the rheme, it is nonetheless the typical approach he takes in thematic analyses (Mathesius 1983: 127).

For Mathesius, it is important to maintain that a singular factor alone cannot account for a text's "objective" realization; instead, a myriad of factors not limited to those listed above is at play at the time of utterance. Reducing a text's degree of expressivity to word order alone would fail the functionalist perspective as well as the inherent complexity that any text may possess. As will become evident in the discussion on subsequent linguists' treatment of thematic constituents, contextual and cotextual factors both complicate and account for deviations in the delineation of thematic elements in a text. Mathesius' initial treatment of theme/rheme boundaries with GIVEN/NEW information status and contextual cues continued to pervade later developments of the thematic paradigm. These developments offered greater delineation of thematic elements yet the lack of consensus to date can be traced back to Mathesius' work presented here.

From the above discussion, Mathesius' work reflects inspiration from Weil's interpretation of discourse message through the concepts of starting point (theme) and enunciation (rheme). Both are reminiscent of Weil's point of departure and the self-same *l'énonciation*. Both linguists' work further revolved largely around the channels involved in communication in terms of the speaker and recipient, although Mathesius formally introduced the topic of context into the communicative equation as a functionalist parameter. Finally, word order, either subjective or objective, was pivotal for understanding how speech users realize and comprehend text. Modern shifts in linguistic approaches are evident in Mathesius' functionalist analysis of word order in discourse and how this is reinforced through information status by means of the theme and rheme.

2.3. Firbas

A fellow member of the Prague School of Linguistics and largely influenced by Mathesius' work, Firbas contributed to thematic theory through the so-called information system and its related communicative dynamism model. In accordance with fellow Prague School scholars, Firbas utilized this model to determine both thematic status and information status. He defines communicative dynamism as "the relative extent to which the unit contributes towards the development of the communication within the communicative field" (Firbas 1996: 221-222). For Firbas, constituents within a text which reflect lower degrees of communicative dynamism are the theme; in turn, those with higher degrees of communicative dynamism are the rheme. Where the issue becomes more complicated is with the GIVEN–NEW dichotomy. Firbas and other Prague School linguists ultimately forego the use of GIVEN and NEW in favor of communicative dynamism. While similarities across both systems exist, stark and unique differences due to communicative dynamism are at play in Firbas' conceptualization of the thematic paradigm. For that reason, what communicative dynamism is exactly, how it is determined and how it affects the delineation of thematic and rhematic constituents will form the discussion of Firbas' work in this section.

Any meaning-carrying constituent within a sentence may possess information, which is to be understood as factual, emotional or attitudinal content with a corresponding degree of communicative dynamism (Firbas 1992: 8). To ascertain the degree of communicative dynamism and its distribution within the sentence, three primary factors are necessary: the

principle of linearity or linear modification, the associated context of the utterance, and the semantics of the utterance.[5] Although all three factors contribute to communicative dynamism distribution, certain factors may take precedence. These factors work in parallel and combinatorially aid in determining how communicative dynamism is distributed amongst the sentence constituents.

Considering the principle of linearity first, words towards the end of the sentence tend to have greater communicative dynamism than those at the beginning of the sentence (Quirk et al. 1985: 1391-1392). Within the system of these three factors, the principle of linearity contributes the least to the communicative dynamism of a text and its constituent parts (Firbas 1992: 9). Consider the sentences with unmarked[6] themes and rhemes in (1) and (2):

(1) He lived in London.
(2) He flew to Prague.

The pronoun *he* in both sentences is the least salient element, carries the smallest degree of communicative dynamism and is there the theme (Firbas 1992: 49). The rheme is *in London* in (1) and *to Prague* in (2). The reason for this is their realization at the end of the sentence and thereby their highest degree of communicative dynamism. These two adverbials additionally function as complements and therefore gain even greater communicative dynamism, as will be shown below in the discussion on context dependence. Firbas generally qualified verbs as rhematic due to their context-independence and due to their grammatical congruence with the subject. The Prague School of Linguistics introduced an additional constituent to the theme and rheme, that of the transition, which is reserved for finite verbs and verbal phrases, as treated more formally below.

While the principle of linearity allows allocating thematic and rhematic elements based on their position within the sentence alone, its contribution is greatest only in unmarked predicative and paratactic utterances (Firbas 1992: 8). Otherwise, this principle is employed when first applying broad strokes to the communicative dynamism distribution within a text before context and semantics are considered. Assuming the principle of linearity to be the superior determining factor would undermine the pivotal role context plays in language realization and thereby the core tenets of functional sentence perspective. This can be illustrated by means of the sample sentences from Firbas (1992), which will also serve as a point of analysis once the factor of context is considered.

(3) He went to Prague in order to meet his friend.

Example (3) reflects standard word order in English. Ignoring the factors of context and semantics for the time being,[7] a gradual increase in communicative dynamism is evident when

[5] While there technically is a fourth factor to consider, namely the communicative intention of the speaker expressed phonetically through intonation, it will not be further explored here since the present work focuses on the written mode.
[6] Unmarked themes and rhemes are those that follow standard word order in the language in question. In English, an SVO language, the grammatical subject realized as theme and the grammatical object realized as rheme are considered unmarked. Depending on whether one adheres to the Prague School or Hallidayan approach to theme/rheme demarcation based on word order, the picture becomes more blurred when considering deviations from standard word order.
[7] It must be stressed here that these two factors are being consciously ignored at this point for the purpose of explaining the principle of linearity. The reason is that the dynamic semantic functions and context, explained below, override and therefore explain the exact differences in thematic and rhematic allocation irrespective of word order.

traversing through the sentence constituents from left to right. Thus, *he* possesses the least communicative dynamism in (3) since it appears first. In so doing, it is attributed thematic status. Conversely, *to Prague* and *his friend* possess the greatest communicative dynamism and are the most salient element in the utterance. These thus form the rheme of (3). Since there are two clauses within this utterance, it is able to possess multiples rhemes as well. In fact, the entire second clause *in order to meet his friend* would be considered more rhematic than *to Prague* since it appears in the latter half of the utterance, therefore gaining communicative dynamism by means of the principle of linearity. For that reason, adverbials appearing finally tend to have greater communicative dynamism than those that appear sentence initially.

(4) Yesterday I met an old friend.
(5) I met an old friend yesterday.

This explains the primary difference in the rhemes in (4) and (5): the communicative dynamism of the text constituents can first be determined by the principle of linearity and reinforced through grammatical dependency of the object complement on the verb *meet*. In (4), *an old friend* would be the most salient, have the greatest communicative dynamism and constitute the rheme. The sentence-initial *yesterday* would have a greater communicative dynamism than the grammatical subject *he* since it appears before the pronoun, but less communicative dynamism than the rhematic predicate. In contrast, *yesterday* in (5) first appears to be most salient on account of its sentence-final position. Looking closely at (4) and (5), however, *an old friend* fulfills the grammatical function of object complement. Despite *yesterday* being the final constituent in (5), such temporal adverbials function as background, concomitant information, not obligatory amplifications (Firbas 1992: 49). This reduces the word's relative degree of communicative dynamism when object complements of the verb are present as well. Hence, the object complement in (5), *an old friend,* realized due to the categorial exponents of the verb reflects the greatest communicative dynamism followed by *yesterday*. Here, the semantics – agent-action-goal – override the principle of linearity. Thus, while a significant factor in the determination of communicative dynamism, the principle of linearity cannot sufficiently account for the myriad of realizations possible in language. Instead, it lays the foundation for an initial understanding of how the communicative dynamism may be distributed within an utterance and is then often overridden by context or semantic factors.

Context, according to Firbas, can be defined in two ways: first, it is viewed as the situational environment which aids the speaker in selecting lexicogrammatical structures to realize a text as a reaction to the situation. The system of context functions in parallel to the lexicogrammatical system, however unidirectionally in that the context limits the otherwise innumerable number of textual realizations. The second meaning of context is associated with the information system. The context can either be dependent or independent and its information can either be known or new, retrievable or irretrievable. As outlined by Firbas: [T]here are 2 types of known information that can be conveyed by the sentence in the active communication: (i) information that, though conveying knowledge shared by the interlocutors, must be considered unknown in regard to the immediately relevant communicative step to be taken and in this sense irretrievable from the context; and (ii) information that not only conveys common knowledge shared by the interlocutors, but is fully retrievable from the context even in regards to the immediately relevant communicative step. (Firbas 1992: 22)

Retrievable here means that either the discourse topic has been explicitly mentioned in the text or that the reader can be derive the topic from the general speech situation. Retrievable is thus equivalent to GIVEN information status, such that it establishes foundational discourse topics throughout the text. In written speech, Svoboda found that a discourse element may continue to

qualify as retrievable if instantiated within the previous seven clauses (Svoboda 1981: 88–89). For Jalilifar, this span is even narrower at only three clauses prior (Jalilifar 2009: 96). Hajičová and Vrbová's investigation into retrievability in spoken speech corroborate this finding with even greater emphasis on immediately following stretches of text (1981: 293-294). By re-instantiating the same element, through repetition, paraphrase, proforms or even ellipsis, the constituent attains and retains its retrievable status. In doing so, it establishes context dependence or known information status as the foundation of the message in discourse. Since the element reflects a context-dependent and therefore retrievable status in the utterance, this element possesses the smallest degree of communicative dynamism and becomes the theme (Firbas 1987: 138).

It should be noted that the theme does not need to be realized sentence initially but rather should be determined as a function of communicative dynamism. Whichever element has the least communicative dynamism in a text is to be understood as the theme. While this is most frequently the context-dependent grammatical subject, certain cases such as imperatives illustrate how the theme can be realized as the grammatical object or even the verb (cf. the thematic *it* in *Get it!* or *come* in *Come over here!*). To reiterate, Firbas defines theme in terms of communicative dynamism by means of context dependency and information retrievability: Context-dependent and thereby retrievable constituents that reflect the least communicative dynamism are to be considered thematic.

On the other end of the scale, rhematic elements are those with the highest degrees of communicative dynamism; therefore, constituents which are context-independent or irretrievable. Firbas concentrates his discussion of the rheme as a function of communicative dynamism around the verb and its so-called competitors. In the most basic of sentences comprised merely of a subject and verb, the subject is assumed to be context-dependent, retrievable, of little communicative dynamism, and therefore thematic. Conversely, the context-independent verb reflects the greater communicative dynamism in a two-element utterance such as *She left*, and takes on rhematic status. In more complex sentences, any additional context-independent element beyond the finite verb functions as competitors to the verb, vying over greater degrees of communicative dynamism and rhematic status within a text (Firbas 1992: 41–42).

The reason why context-independent elements other than the verb carry greater communicative dynamism is because of these elements' ability to push forward the communication by adding to, specifying or expounding on informational content within the communication, i.e., the exact function of the rheme. The verb itself may partially contribute to forwarding discourse, but elements such as subject or object complements, temporal or locative adverbials and even adjuncts aid in developing communication further or even completing it. Adverbials in particular play a critical role in communicative development and therefore merit special treatment, as alluded to above in the discussion on the principle of linearity.

Returning to the previous examples, restated in (6) – (8) in the following, the principle of linearity ultimately determined the communicative dynamism distribution of the sentence constituents in (6) and (7). In (8), however, the semantic factor overrode the principle of linearity. Before looking at how semantics is incorporated into a text's communicative dynamism, how context (in)dependence and information retrievability applies to these examples will be explained.

(6) He went to Prague in order to meet his friend.
(7) Yesterday I met an old friend.
(8) I obviously met an old friend yesterday.

According to Firbas, adverbials are able to fulfill three different communicative functions: **specification**, **setting** and **modality** or indefinite time adverbials (Firbas 1992: 49, 77-78). Comparable to the distinction between complement and adjunct in terms of obligatory and optional grammatical realization, specification adverbials hold greater communicative dynamism than setting adverbials due to their grammatical or semantic necessity. In the cases above, *to Prague* in (6) would qualify as a context-independent adverbial of the specification function due to its dependence on the verb *went*, whereas *yesterday* in (7) and (8) is a context-independent adverbial of setting function. As such, specification adverbials offer obligatory and amplificatory information essential to the core of the message, which results in greater communicative dynamism. It should, at this point, be noted that Firbas' delineation of "obligatory amplifying information" is expressed loosely (cf. Firbas 1992: 50–51), such that his argumentation would otherwise fail in (6). The reason is that a sentence such as *He went* would, grammatically, be acceptable in response to the question *He did what?* The point Firbas makes, however, is that the specifying adverbial *to Prague* contributes to the communicative development of the text. It offers critical information that amplifies the core message being conveyed. The corresponding prompt to elicit a response such as (6) would be *Where did he go?, What did he do?* or *Why did he go to Prague?*

Conversely, facultative adverbials of the **setting** type, i.e., adjuncts, offer additional yet concomitant information as in (7) and (8). Their contribution does indeed gain greater communicative dynamism than that of the verb since they act as a competitor to the verb. However, their degree of communicative dynamism remains subordinate to subject or object complements and specification adverbials. Should an adverbial be realized as context-dependent, then the adverbial invariably takes on a setting function irrespective of sentence position. Finally, it is important to remember that context-independent adverbials of either type reflect lesser communicative dynamism when realized sentence initially as opposed to finally due to the principle of linearity. Typically, however, sentence-initial adverbials qualify as settings, those in final or near-final position as specifications.

Finally, adverbials of the **modal** type are not considered to be a competitor of the verb due their facultative nature and lacking amplification of the core of the message, i.e., they contribute to the foundation, not the development, of the message (Firbas 1992: 50, 77). In (8), the modal adverbial *obviously* is considered context-independent since it offers additional but not obligatory information. Together with the finite verb, modal adverbials constitute the transition, which is the boundary between the theme and rheme unique to the Prague School's model of thematic theory. While modal adverbials do not amplify the core of the message, they do possess a higher degree of communicative dynamism than the finite verb. Therefore, the finite verb *met* in (8) would have the lower degree of communicative dynamism than that of the modal adverbial *obviously*.

The examples above illustrate how context and syntactic realization affect the retrievability and thereby the communicative dynamism of discourse topics within communication. With the calculated communicative dynamism, the text's thematic, transition and rhematic elements can be delineated in accordance with their grammatical and semantic relationship to the verb. The communicative dynamism of a message can finally be reinforced through the underlying semantic system that governs the realization of a text.

The semantics of a message, more accurately defined as dynamic semantic functions, build upon the informational actively accessible from the immediate context and the constituents of the utterance. Within this framework, Firbas presents two scales, the **Presentation Scale** and the **Quality Scale**, which are defined as follows:

a) SETTING – PRESENTATION OF PHENOMENON – PHENOMENON PRESENTED
b) SETTING – BEARER OF QUALITY – ASCRIPTION OF QUALITY – QUALITY – SPECIFICATION – FURTHER SPECIFICATION

and can even be combined into a single **Combined Scale**:

c) SETTING – PRESENTATION – PHENOMENON PRESENTED – BEARER – QUALITY – SPECIFICATION – FURTHER SPECIFICATION

The scales' elements are arranged according to a rising degree of communicative dynamism and can be applied to any text as a final means of determining its communicative dynamism distribution. These scales reflect the semantic information of a text irrespective of word order, and each utterance only reflects one of these scales. This is because the scales are to be understood as an **interpretative arrangement** of a text's constituents and communicative dynamism distribution. The interpretive arrangement stands in contrast with the **actual linear arrangement** of the text, i.e., the text that was syntagmatically realized. By means of both arrangements, it is possible to examine whether the communicative dynamism distribution expressed through word order (principle of linearity) corroborates the communicative dynamism distribution expressed through the semantics (dynamic semantic functions) of the text.

To illustrate how dynamic semantic functions affect communicative dynamism, the example sentences in Figure 2-2 to Figure 2-4 as inspired by work from Adam (2013) will be considered. These will also indicate differences across the three scales and how they can be used to determine the theme, transition and rheme of a text.

	Theme	Transition	Rheme
Presentation Scale	SETTING	PRESENTATION OF PHENOMENON	PHENOMENON PRESENTED
Communicative Dynamism (CD)	LOWEST CD →		HIGHEST CD
Text	*Over the hill*	*came*	*a cart and ox.*

Figure 2-2: Use of the Presentation Scale and communicative dynamism to determine the elements of theme, transition and rheme.

The **Presentation Scale** exemplified in Figure 2-2 is the simplest of the three scales and is applicable in cases where a verb of existence or appearance is employed, such as *come, appear, occur, turn up, happen, arrive,* or *come up* (Firbas 1992: 60). Verbs are represented by the PRESENTATION OF PHENOMENON in the scale and constitute the second most salient element in the utterance. The SETTING is typically realized by temporal or spatial adverbials, like *over the hill* in Figure 2-2, and reflect the least communicative dynamism within the Presentation Scale. For that reason, the element(s) attributed to the dynamic semantic function of SETTING are thematic. The grammatical subject *a cart and ox* form the PHENOMENON PRESENTED within the Presentation Scale. On account of the principle of linearity, these sentence-final constituents have the highest communicative dynamism and become the rheme. This is further reinforced by the indefinite article, which nearly always reflects the rheme (McCabe 1999:172). The elements *a cart and ox* forming the rheme may be unexpected since grammatical subjects tend

to be thematic. This is indeed the case when texts are formulated whose dynamic semantic functions fall into the Quality Scale. However, since the grammatical subject appears sentence finally and the prepositional phrase *over the hill* has a lower communicative dynamism through its SETTING qualification, *a cart and ox* must become rhematic.

The Presentation Scale is employed most prototypically at the beginning of a communication, whether written or spoken. One might think of the beginning of fairy tales or children's stories which conventionally begin with *Once upon a time there was a princess*. In instances such as these, the reader is presented with a princess whose existence is set relative to a time long ago. Here, the actual linear arrangement coincides with the interpretive arrangement, whereby the order in which the sentence constituents appear is an exact reflection of their dynamic semantic function and communicative dynamism distribution.

	Theme		Transition	Rheme		
Quality Scale	SETTING	BEARER OF QUALITY	ASCRIPTION OF QUALITY	QUALITY	SPECIFICATION	FURTHER SPECIFICATION
Communicative Dynamism (CD)	LOWEST CD →					HIGHEST CD
Text	*In the past*	*he*	*felt*	*jealous*	*of his friend's bike*	*a lot*

Figure 2-3: Use of the Quality Scale and communicative dynamism to determine thematic, transition and rhematic constituents.

The second scale, the **Quality Scale**, is not only more frequent than the Presentation Scale but also more complicated, not least because it is comprised of more dynamic semantic functions. However, not all elements of a given scale need to be instantiated. In Figure 2-3, all functions of the Quality Scale are accounted for and, for simplicity's sake, coincide directly with the sentence constituents. Therefore, both the interpretive and actual linear arrangement are the same.

Similar to the Presentation Scale, the Quality Scale starts with the SETTING dynamic semantic function. This indicates the element with the least communicative dynamism and provides clear delineation of (one of) the thematic element(s). As alluded to previously, the grammatical subject, which takes on the function of BEARER OF QUALITY, is always thematic despite it reflecting a comparatively higher degree of communicative dynamism than the SETTING element. Next, the ASCRIPTION OF QUALITY can only be realized by copular verbs, *felt* in Figure 2-3, which are completed with an adjective or noun complement in the QUALITY slot, here *jealous*. Should an utterance not make use of a copular verb, then the ASCRIPTION OF QUALITY function remains unfulfilled and the verb in question then fills the QUALITY slot.

These two dynamic semantic functions are of particular interest as they represent the so-called transition element unique to Prague School linguists' interpretation of the thematic paradigm. This element functionally and positionally acts as the boundary between thematic and rhematic elements, and will be discussed below once the discussion on semantics has concluded.

The final two elements of the Quality Scale, *of his friend's bike* and *a lot*, correspond to the SPECIFICATION and FURTHER SPECIFICATION functions, respectively, the latter of which is optional. These two elements reflect the highest degree of communicative dynamism and are therefore the rhematic elements of the text. The facultative FURTHER SPECIFICATION has even more communicative dynamism than that of SPECIFICATION. The Quality Scale is then the default scale used to determine communicative dynamism when a text contains no verbs of existence or appearance. Hence, this scale is used most frequently amongst the three provided.

Finally, the **Combined Scale** (cf. Figure 2-4), as the name suggests, is a combination of both the Presentation and the Quality Scale. Here, "the distributional field telescopes the [PHENOMENON]-function and the [BEARER]-function into the subject" (Firbas 1992: 67); in other words, the syntactic and semantic functions of both scales are instantiated simultaneously and to be interpreted combinatorially. At opposite ends of the scale, the SETTING and SPECIFICATION functions form the theme and rheme, respectively. This is due to the SETTING having the least communicative dynamism and (FURTHER) SPECIFICATION having the greatest communicative dynamism. Additionally, the theme is also expressed by the grammatical subject, which corresponds to both dynamic semantic functions PHENOMENON PRESENTED and BEARER OF QUALITY. Similar to the example sentence in Figure 2-4, the theme is again typically realized through an indefinite noun phrase, here *a new city*. Next, the QUALITY function is fulfilled by the verb *was built*, which can also be a verb of appearance or existence. Similar to the Quality Scale, the verb in the Combined Scale represents the median of the sentence by reflecting communicative dynamism greater than that of the theme but less than the rheme. The finite verb phrase again constitutes the transition element in terms of thematicity. Finally, the prepositional phrase *of gold* takes on the SPECIFICATION function and receives greatest communicative dynamism on account of its grammatical dependency on the verb and realization at the end of the sentence.

	Theme		Transition	Rheme
Combined Scale	SETTING	BEARER OF QUALITY	QUALITY	SPECIFICATION
Communicative Dynamism (CD)	LOWEST CD	→		HIGHEST CD
Text	*Moons ago*	*a new city*	*was built*	*of gold.*

Figure 2-4: Use of the Combined Scale and communicative dynamism to determine the elements of theme, transition and rheme.

The Combined Scale is commonly employed when complex texts are realized with verbs of appearance or existence. Since the Presentation Scale lacks the SPECIFICATION function, the Combined Scale allows for an expanded analysis. It should be noted that Chamonikolasová & Adam (2005) found evidence against the necessity of the Combined Scale, instead arguing in favor of the Presentation Scale and Quality Scale alone. That being said, all three scales can be used in determining text constituents' communicative dynamism in the Prague School approach to the thematic paradigm.

Together with these three scales of dynamic semantic function, a complete picture of how communicative dynamism can be calculated in written text has emerged. Due to their complexity, the purpose these three factors have in communicative dynamism and subsequent determination of thematic elements will be recapitulated. With the principle of linearity, the simplest of the three factors, an initial and general understanding of the communicative dynamism distribution can be gleaned. While most insightful in conventional SVO sentences, this first factor is commonly overridden by the effect that context and the semantics have on an element's communicative dynamism.

Context, in the pragmatic sentence, facilitates the lexicogrammatical selection process that shapes the foundation and core of the message to be realized. In its narrower, Firbasian meaning, context is also understood in terms of information retrievability. Information that interlocuters have no access to within the discursive context qualifies as irretrievable and therefore context independent. This can either be due to the absence of explicit realization within the discourse or due to having exceeded the limitations of the information processability (up to seven clauses prior). Context-independent elements invariably reflect higher degrees of communicative

dynamism, regardless of syntactic position, and stand in juxtaposition to context-dependent elements. These are constituents that have been previously mentioned in the discourse and are therefore retrievable information for the interlocutor. Finally, context-dependent elements reflect the lowest communicative dynamism in an utterance. They form the foundation of a message through their repeated, paraphrased, elliptical or coreferential use. Hence, within the functional sentence perspective framework, context-independent elements are responsible for developing discourse through their foregrounded status and special treatment in the dissection of all communicative development. Not only is context a pivotal and foundational principle of functional sentence perspective, it is also a governing aspect in how communication evolves and is understood.

Serving as both a parallel and complementary system to context, dynamic semantic functions prove to be a complex yet discerning means for communicative dynamism distribution on the basis of the principle of linearity and context. This system provides the necessary information on how the individual constituents of a text contribute to the ebb and flow of communicative dynamism. The degree of communicative dynamism vis-à-vis dynamic semantic functions is determined irrespective of word order yet as a factor of relative salience. The Presentation, Quality and Combined Scales further account for the intricacies of language expression by allocating communicative dynamism to text constituents and thereby the respective scales' functions. The dynamic semantic functions are lastly able to confirm and thereby reinforce the contribution that both the principle of linearity and context provide.

Such a tripartite approach to determining the communicative dynamism distribution of a text represents the pioneering work Firbas achieved. It furthered the theoretical and practical understanding of the theme-rheme dichotomy in text and facilitated the conceptualization of the transition element unique to the Prague School.

Whereas Mathesius' interpretation of the thematic paradigm accounted for theme and rheme alone, Firbas' model enabled text to be dissected into a minimum of two and up to six parts.[8] Contrary to thematic theories prior to Firbas, a minimally realized text is not comprised of theme and rheme alone but rather of the transition and rheme proper. Hence, texts are able to be realized without a theme, albeit scarcely.

	Theme		Transition		Rheme	
Thematicity	THEME PROPER	DIATHEME	TRANSITION PROPER	TRANSITION	RHEME	RHEME PROPER
Quality Scale	SETTING	BEARER	QUALITY		SPECIFICATION	FURTHER SPECIFICATION
CD	LOWEST CD →					HIGHEST CD
Text	*In the coming years*	*Mr. Grave*	*will clearly*	*have*	*greater responsibility*	*for others.*

Figure 2-5: Further delineation of thematic, transition and rhematic elements through the introduction of additional proper *elements according to Firbas and with the help of communicative dynamism (CD).*

[8] Depending on methodological approach, some linguists have added even more thematic, transition and rhematic elements, such as transition proper oriented, transition to the exclusion of transition proper and transition proper oriented elements, question focus anticipator and negative focus anticipator (cf. Firbas 1992). While the inclusion of these elements may allow for greater degree of fine-tuned analysis, the present discussion focuses solely on Firbas and Svoda's (1986 and 1981, respectively) tripartite approach with thematic, transition and rhematic elements alone.

The addition of the term *proper* to the transition and rheme represents yet another level of abstraction in the analysis of a text's thematic constituency: The theme is further delineated by the theme proper and diatheme, the transition by transition proper and transition, and the rheme by rheme proper and rheme. A fully realized text is illustrated in Figure 2-5 with its respective communicative dynamism and thematicity:

The theme proper, realized here as a temporal adverbial fulfilling the dynamic semantic function of SETTING, possesses the least communicative dynamism within the utterance for contextual and semantic reasons described above. As such, the theme proper is most frequently realized by adverbials of the SETTING function, conjunctives or entire subordinate clauses. The diatheme became a subsequent part of Firbas' conceptualization of the thematic paradigm following work by Svoboda (1981). According to Svoboda, the diatheme can be defined as:

a) The thematic element carrying the highest degree of communicative dynamism (Svoboda 1981: 5).
b) The temporary center of the scene, the newly introduced or just chosen quality bearer (Svoboda 1981: 42).
c) What Mathesius originally referred to as "the centre of the theme." (Svoboda 1981: 5).

Following these characterizations, both a) and b) are at play in the example sentence from Figure 2-5. The subject *Mr. Grave* functions as QUALITY BEARER in the Quality Scale, thereby allocating it greater communicative dynamism than the theme proper's SETTING function. Definition c) is more of a characterization and less operationalizable than a) or b). Hence, using it as a justifiable explanation for attributing diathematic status to a sentence constituent should be done with caution. Should a text only have one thematic element, then the constituent belongs to both the theme proper and diatheme (Svoboda 1981: 6).

The next primary sphere of a text is that of the transition, which is invariably determined by the verb and its so-called temporal and modal exponents (TMEs). These are first and foremost responsible for "[serving] as a link in as a boundary between the foundation of the message and its core" (Firbas 1992: 71). These functionally represent the delineation between thematic and rhematic elements in the sentence. They initially reflect the greatest degree of communicative dynamism in a text if no other competitors are present, i.e., no text follows the finite verb. Verbs that function as transition proper are typically modals or auxiliaries and hence indicate modality, temporality, aspectuality, tense and mood. If an adverbial of indefinite time, such as *generally, usually,* or *sometimes* is employed, then it qualifies as the transition proper. Conversely, copular and full or lexical verbs are reserved for the role of transition. Categorial verbs that require complements are of particular interest here as they functionally reflect the transitional character into the rheme.

The rheme corresponds to the dynamic semantic function of (FURTHER) SPECIFICATION and entails the communicative purpose of the message; that is, the propositional content that progresses the information from the foundation of the message. The element that reflects the highest degree of communicative dynamism is the rheme proper, which, in Figure 2-5, is the complement *for others*. This prepositional phrase is linked to the sentence constituent with the second highest amount of communicative dynamism *greater responsibility*. The phrase *for others* has the highest communicative dynamism due to both the principle of linearity and its dependency on *greater responsibility* as a prepositional complement.

The final point to mention in Firbas' contribution to the thematic paradigm is that of hyperthematicity. Firbas explains this concept within the context of givenness or contextual relevance: here, not only single utterances but entire texts themselves have themes of their own, which constitute a text's hypertheme(s) (Firbas 1992: 109–110). In non-linguistic terms, a hypertheme is the general topic, gist or thread of a text that can span a paragraph, chapter or even an entire book. The themes within a particular text boundary (a paragraph, a chapter, etc.) are thus set in relation to the givenness provided by the hypertheme. The themes of each sentence form the foundation of the discourse and thereby the collective anchoring across a set of sentences. This comprehensive collection of themes throughout a text contributes to its overall hypertheme and reinforce the themes' GIVEN information status. By associating the contribution of individual themes to a greater hypertheme, Firbas laid the foundation for a fellow Prague School linguist, Daneš, to conceptualize the system of thematic progression. This came to define how text and its propositional content can be analyzed in terms of its syntactic realization from sentence to sentence on the basis of thematic elements' informational status GIVEN and NEW.

Building off the work of Mathesius, Firbas furthered the thematic paradigm by conceptualizing novel systems that facilitate the structural and propositional analysis of a text. Within the framework of functional sentence perspective, the systems of syntactic realization, semantics, and contextual information all function in parallel to reveal a text's communicative dynamism and thereby communicative purpose. Whereas the principle of linearity minimally affects the overall degree of communicative dynamism a text may have, contextual (in)dependence proves to be a decisive factor. It unequivocally aids in determining both the thematic foundation of a text and rhematic core of the discourse message. The semantics of the individual sentence constituents, particularly that of verbs as well as temporal and modal exponents, provide the additional layer of interpretive systematicity necessary in the final ascription of thematic, transition and rhematic status. Through a tripartite approach to analyzing the communicative purpose behind text, Firbas operationalized the dissection of a text's thematic constituents and how information status is attained by means of communicative dynamism.

2.4. Daneš

The next Prague School linguist to have forwarded the theoretical framework of thematic theory was Daneš, who first formalized the analysis of discourse through so-called thematic progression (Daneš 1974). This model leverages themes and rhemes within text to trace how the communicative message of each sentence develops throughout the entirety of the discourse. Thematic progression and his approach to thematic theory are only briefly introduced in the present section. A more comprehensive treatment of thematic progression is given in Chapter 3.1, where thematic theory and thematic progression are elucidated as a model for the analysis of discourse structure and development.

In line with previous Prague School linguists, Daneš viewed the theme as the point of departure of a text, however one of deliberate and conscious choice (Downing 2001: 22). From Daneš's point of view, previous situational and linguistic contexts within a discourse aid the speaker in the selection of themes, since the interlocutor would otherwise be overwhelmed by the sheer possibilities of linguistic expression. In terms of determining which sentence constituents are thematic, he followed subscribed to the Firbasian approach. Aside from his explication of hypertheme by means of thematic progression, Daneš developed the GIVEN-NEW dichotomy even further as was spurred on by Halliday's interpretation.

As outlined in Daneš (1974: 109–110), givenness is defined according to the following criteria:

1. GIVEN information is derivable or recoverable from the context, situation and the common knowledge of the speaker and listener.
2. Givenness, similar to [communicative dynamism], is not absolutely binary but graded.
3. The degree of givenness depends on the proximal recoverability of the element in question.
4. Givenness in terms of contextual information may be provided in terms of an utterance's semantics, cohesive devices or pragmatic information, such as implicature, inference or contextual delicacy.
5. Whereas givenness has received ample treatment in operationalizability, newness is to be understood in terms of communicative relevance according to the speaker's intentions.

What should stand out the most from these explanations of givenness is their similarity to Firbas' understanding of givenness. Particularly points (1) – (3) are ones that Firbas addressed directly and were touched on in Chapter 2.3. Upon closer examination, however, these definitions show that Daneš contextualized recoverable information around presuppositions from the perspective of the text recipient. The text producer might presume a great deal of what the recipient may have access to in terms of common knowledge. Otherwise, regurgitated information in overtly or implicitly repeated (e.g., paraphrased) form may more clearly signal to the listener the givenness of the information under discussion (cf. point (4)).

Additionally, felicity conditions may also be at play, where pragmatics come to affect the realization of a text. False assumption of presumed knowledge can ultimately result in infelicitous utterances or failed illocutionary acts (Austin et al. 1975; Hawes 2015: 94). The pragmatic, and thereby contextual and/or illocutionary, effect is further motivated by the speaker's intentions and what a speaker intends to be understood as NEW (cf. point (5) above). Note that this stands in contrast to Mathesius and Firbas' interpretation of NEW. The two defined this as information irrecoverable from the context. In contrast, Daneš placed greater value on the *intention* behind what should be received as NEW within a text (1974: 109-110). In spoken speech, this can more readily be traced via prosody (Adam 2013: 16). However, in written text, speech markers indicating intention are less overt or simply absent, which makes determining GIVEN and NEW via intention much more difficult if not entirely subjective (Downing 2001: 21).

Be that as it may, certain aspects of written speech can qualify as NEW upon first identifying GIVEN information following the assumptions (1) – (4) above. An important point to mention here is how Daneš associates GIVEN and NEW with the theme and rheme. Like many other Prague School linguists, Daneš takes what is called the 'combining approach', whereby GIVEN is thematic and NEW is rhematic (Fries 1981: 36). This also aids in determining the boundary between GIVEN and NEW since there are clearer delineations as to what qualifies as thematic and rhematic in written text.

Aside from his understanding of GIVEN vs. NEW, Daneš's greatest contribution is how he relates themes to the unfolding of a text's overall structure. He first achieved this through a continuation of Firbas' concept of hypertheme, such that the themes of concatenated text belong to overarching themes that functionally resemble hypernyms: Just as one would expect sentences within a single paragraph to informationally be related to a greater topic, or hypertheme, these paragraphs should also contribute to the greater theme of a chapter. This, in turn, contributes to the hypertheme of the entire text. In associating the themes with hyperthemes, Daneš took defining steps towards connecting discourse at the micro level with

that at the macro level of the text (Dubois 1987: 108). Thus, the dynamicism present in individual, static sentences is brought to the forefront through the inner connexity of texts. This manifests through *"*the choice and ordering of utterance themes, their mutual concatenation and hierarchy, as well as their relation to the hyperthemes of the superior text unit (such as paragraph, chapter, etc.), to the whole text, and to the situation" *(*Daneš 1974: 114).

This quote illustrates Daneš's conceptualization of what came to be called thematic progression as a model for the network of thematic connections (or connexity) that any text has.[9] Both the initial and subsequent realizations of a theme based on situational and contextual parameters prime the choice of future themes. These are further primed by the propositional content in rhemes. However, in terms of actual structural contribution, themes take precedence according to Daneš since these form the foundation of the discourse message.

The theme's precedence becomes evident in the four types of thematic progression that Daneš postulated: simple linear progression, constant continuous progression, themes derived from a hypertheme and split rheme progression. Simple linear progression is instantiated when rhematic elements from one sentence are realized as the theme of the subsequent sentence. Constant continuous progression is the re-instantiation of the same theme across sentences. A theme derived from a hypertheme is the instantiation of a theme from a superordinate discourse topic representative of the entire text (cf. Chapter 3.1 for sample texts of these patterns and how they contribute to the development of discourse in practice). Finally, split rheme progression is the re-instantiation of listed discourse topics in the rheme of one sentence as themes in subsequent sentences. Functionally, these progression patterns reveal the overall thematic structure of a text and how it is developed through thematic selection. While rhematic elements are considered in thematic progression, they are always contextualized in terms of how they become subsequent themes. For that reason and due to the theme's function as the foundation of the discourse message, Daneš's progression models were termed thematic and not rhematic progression despite rhematic progression patterns emerging later (cf. Leong 2005: 712; Hawes 2001; Li 2009).

Overall, Daneš's contribution to the thematic paradigm was the further refinement of the GIVEN-NEW dichotomy and the formalization of themes and rhemes' contribution to discourse development through thematic progression. Whereas Weil, Mathesius and Firbas implied a sense of progression in communicative content from the foundation to the core of the message, their treatment of theme and rheme primarily centered around the analysis of individual clauses alone. Daneš's thematic models conversely embedded dynamicism into discourse analysis that the theme and rheme afford during the realization and unfolding of discourse. His theoretical contribution to text linguistics thereby represents a turning point in how the thematic paradigm is understood and employed in text analysis. The ramifications that Daneš's thematic progression models have had on the methodological approach to discourse analysis are evident in the continued refinements to his original models that have shaped contemporary research on thematic progression and thematic theory.

[9] Daneš argues that text connexity is not a binary parameter but more so rests on a cline; therefore, texts may exhibit varying degrees of text connexity depending on context and purpose (Downing 2001: 21).

2.5. Halliday

The final linguist to be discussed in the context of the further conceptualization of thematic theory is Halliday, the founder of systemic functional grammar. This approach to language as a social semiotic system considers the network of choices that language users make to realize discourse and is informed by Halliday's model of transitivity (Halliday 1967: 37-38). The core tenets of transitivity will be outlined in the following specifically with respect to theme and rheme alone. Therefore, while a wealth of information on the system of transitivity as a language model exists, the present discussion will only focus on its use for the delineation of theme and rheme from a systemic functional grammar perspective.

Systemic functional grammar as a linguistic framework follows a similar functional approach to language as the Prague School: it "involves explaining why a given phenomenon occurs by showing what constitutes its contribution to the text in question or to a larger system of which it is itself a subsystem" (Downing 2001: 16). Text is "an instance of the linguistic system […] operating in a context of a situation" (Halliday & Matthiessen 2014: 46-47). This then applies to either the written or spoken mode. Realizations are enabled via social interaction and as "objective options that the linguistic systems allow in a specific situation against the background of a particular culture" (Halliday 1974: 48–49). For Halliday, the Firthian concept of a system possessing a certain set of features underlies all linguistic phenomena. Features that a language user may choose from within the system are mood, polarity, modality and number, to name a few. Further, the role of a 'specific situation', i.e., context, ultimately affects the paradigmatic options available to the language user when syntagmatically producing language. A context-driven systematicity of choice and ultimate realizational patterns thus allows for infinitely possible meaning potentials despite a limited feature inventory.

What further underlies the core of these features and the language system in which they are embedded are Halliday's so-called metafunctions. These are divided into experiential (also known as ideational), interpersonal and textual, and aid in an understanding of theme and rheme according to Halliday. These three metafunctions are summarized in Table 2-1 with the grammatical function they can fulfill.

Metafunction	Definition	Function
Interpersonal	expresses speech roles; provides interactants involved in dialogue with the resources for enacting speech functions through the grammar of the clause	modal adjunct, vocative, finite verbal operator
Experiential	expresses the speaker's inner and external world; functions as a participant, circumstance or process	predicator, subject, complement, adjunct
Textual	relates utterances to the situation and binds the text together; functions as a cohesive tie across sentences	continuative, conjunctive, adjunctive

Table 2-1: Definition and grammatical function of Halliday's metafunctions (edited from Halliday & Matthiessen 2014: 61).

These metafunctions manifest through the sentence constituents realized in a text, which ultimately affect where both the theme and rheme lie within a clause. Halliday initially took inspiration from Mathesius' definition of theme as a text's point of departure: "The theme is

what is being talked about, the point of departure for the clause as a message...” (Halliday 1967: 212). In subsequent decades, Halliday further refined his definition:

- “The English clause consists of a ‘theme’ and a ‘rheme’ […] The theme of the clause is the element which, in English, is put in first position” (Halliday 1970: 161);
- “The theme is the element which serves as the point of departure of the message; it is that with which the clause is concerned” (Halliday 1985: 39);
- “That which is placed in the initial position, is given information serving as ‘the point of departure’ for the clause or which locates and orients the clause within its context” (Halliday & Matthiessen 2014: 89).

The first point to recognize in Halliday’s definitions is his explicit mention of the location of the theme in English. Halliday argues that sentence-initial elements in writing are inherently thematic, at least for English. Later, Halliday incorporated his system of metafunctions into the definition, such that “the theme of the clause is the first group or phrase that has some function in the experiential structure of the clause” (Halliday & Matthiessen 2014: 91). Halliday thus defines the experiential constituent as the **topical theme** of the sentence. Either participants or circumstantial adjuncts of the verb as outlined in his system of transitivity qualify as topical themes (Halliday & Matthiessen 2014: 156). In the sentence

(1) He is there.

the topical theme *he* is the participant of the verb as the first (and only) experiential constituent; the remainder of the sentence and thereby predicate *is there* then forms the rheme. Conversely, in the sentence

(2) In the morning, he is there.

only *in the morning* is the topical theme since it is the circumstantial adjunct of the verb that fulfills the experiential metafunction. *He* in (2) is then relegated to the rheme, which, in its entirety, equates to *he is there*. Since the topical theme, *he* in (1) and *in the morning* in (2), entails the experiential metafunction of the clause, it is the foundation upon which the remainder of the text, i.e., the rheme, rests.

Aside from the obligatory topical theme, a standard English sentence may also contain optional **interpersonal themes** and **textual themes**. Should all three types of themes be realized in a single text, the typical ordering of these elements is textual–interpersonal–topical (Leong 2015: 292). Due to their facultative nature, textual and interpersonal themes either function as cohesive ties or present propositional, contextualizing or modal content, respectively (Leong 2005: 704).

Textual themes establish cohesion and, more often than not, coherence across clauses. They thus serve the function of developing the logic and structure of a text. In terms of their formal characteristics, textual themes can be continuative adjuncts, conjunctive adjuncts or conjunctions. Continuative adjuncts are typically interjections (*oh, yes, well*), conjunctive adjuncts are structural cohesive devices (*so, further, besides*), and conjunctions are coordinating or subordinating conjunctions and relative pronouns (*and, when, which*). While realized most readily at the beginning of a clause, adjuncts and conjuncts can also appear sentence medially to establish cohesion (Eggins 2004; Gerot & Wignell 1994; Zahra et al. 2021).

Interpersonal themes, conversely, express a writer's stance, attitude, evaluation, modality or degree of certainty toward a proposition. Their class belongs to that of adverbial adjuncts functioning as "modal/comment, […], finite/verbal operator in yes/no interrogative, mood, polarity or any combination of vocatives or personal names" (Yuned 2016: 201). Aside from modal adjuncts (e.g., *clearly*, *surprisingly*), vocatives (e.g., *Brian!*) and mood polarity (e.g., modal verbs such as *could*, *may*, *must*), interpersonal themes are uniquely realized in polar interrogatives as the finite verb (e.g., *did* in *Did you know?*), which otherwise belongs to the rheme in standard, declarative sentences. Through the realization of interpersonal themes, the language user establishes and conveys their relationship to the text recipient. This is contextualized within the ideational and textual realization of the message, i.e., the lexicogrammatical choices selected to create meaning according to the discourse context (Forey 2002: 47-48).

All three theme types with their corresponding metafunctions are illustrated in Figure 2-6 from Halliday & Matthiessen (2014: 107). The noun phrase *the best idea* is the participant of the verb *wouldn't…be* and the grammatical subject, which corresponds to the experiential metafunction in the statement. For that reason, it becomes the topical theme of the interrogative. Since textual and interpersonal themes cannot fulfill the experiential metafunction, they cannot qualify as the topical theme (Jalilifar 2009: 86).

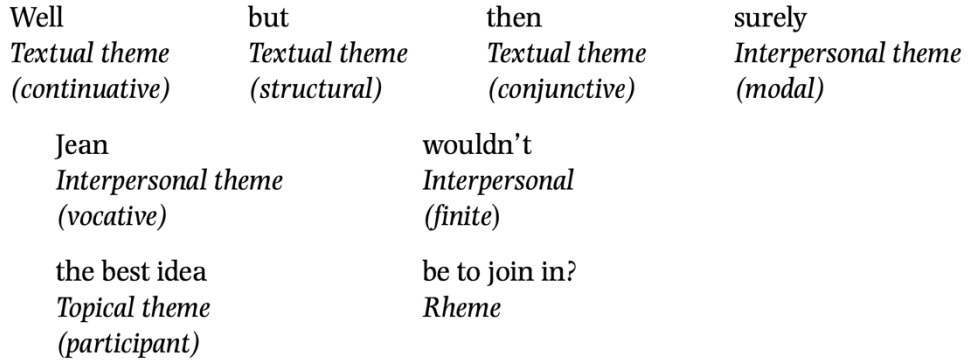


Figure 2-6: Thematic metafunctions – interpersonal, textual and topical – in practice as put forward by Halliday & Matthiessen (2014: 107).

The adjunct adverbials functioning as textual themes serve as cohesive ties (*well, but, then*) and the modal stance markers fulfill the interpersonal metafunction as interpersonal themes (*surely, Jean, wouldn't*). The fact that the textual and interpersonal themes are fronted and not realized within the rheme indicates their contribution to the message's point of departure. To reiterate, the textual metafunction "[construes] experience and enacts interpersonal relations […] to build up sequences of discourse, organizing the discursive flow, and creating cohesion" (Halliday & Matthiessen 2014: 30-31). In other words, it is through the textual metafunction that expression of experience is organized textually within discourse. It provides direction, cohesion and structure. The interpersonal metafunction, on the other hand, indicates the speaker's stance towards a message's proposition: confirmation or refutation thereof, compliance or refusal, negotiation or acceptance. The individual and their emotional or psychological outlook towards a message's proposition are therefore encoded in the interpersonal (Halliday & Matthiessen 2014: 111).

As the number of metafunctions increase in a text, so, too, does the informational weight and complexity. Here, informational weight can be seen as the degree of contextualization and structuring through fronted elements. In other words, multiple instances of textual, interpersonal and even circumstantial (topical) themes realized at the beginning of a sentence

increase their informational weight through the discourse functions they fulfill. An increase in (thematic) realizational complexity is thus directly proportional to informational weight and captured by the sentence constituents' markedness.[10]

Markedness is of particular importance within the thematic paradigm as it accounts for a deviation from or expansion of the standard SVO sentence structure. A realization of both an unmarked and a marked theme is exemplified in the two sample sentences in Figure 2-7. In the first sentence, the grammatical subject *I* is realized without any fronted adjuncts, complements or conjunctions. The grammatical subject, and thereby topical theme, *I* is realized first and follows standard SVO structure. As such, it is given unmarked thematic status. In the second sentence, however, the prepositional phrase *in the morning commute* begins the statement and functions as a circumstantial adjunct. Since it is not the grammatical subject but fulfills the experiential metafunction, it is afforded marked status.

UNMARKED THEME	RHEME	MARKED THEME	RHEME
I	*woke up at 4.*	*In the morning commute,*	*there was hardly any traffic.*

Figure 2-7: Thematic constituents functioning as the grammatical subject such as I *in the first sentence are considered unmarked. Conversely, sentence-initial adjuncts like* in the morning commute *appearing before the grammatical subject are marked as in the second sentence.*

So long as a sentence follows standard SVO structure, whereby no constituents are realized before the grammatical subject, the topical theme is unmarked. Any sentence constituents realized before the grammatical subject, however, become marked and the grammatical subject is then relegated to the rheme. While Halliday outlines the dichotomy of marked/unmarked themes, he does not assume the same dichotomy for rhemes. This, according to Taglicht, weakens Halliday's model (see Taglicht 1984: 23–24). Although not part of the Hallidayan model, it is worth mentioning what constitutes a marked rheme since Taglicht is not alone in advocating for marked/unmarked rhemes (see also Dubois 1987).

Marked rhemes can be substantiated by their divergent word order, just as a deviation from the standard SVO structure merits marked themes. Constituents appearing to the left of the unmarked rheme position, per Dubois (1987), or as shifted elements, per Taglicht (1984), qualify as marked rhemes. This begs the question of whether the theme, marked or otherwise, is the position to the left of the rheme. Taglicht (1984: 23–24) makes use of so-called a) end-shifted subjects, b) end-shifted predicative elements and c) partitioning elements to explain marked rhemes, which are shown in bold in (3) – (5), respectively:

(3) Into the room came **a strange man**.
(4) There were **in this article** a number of interesting points.
(5) They are returning, **however**, to England.

Firstly, it should be noted that Halliday would consider all rhemes in (3) – (5) as unmarked. For Taglicht, end-shifted subjects such as *a strange man* in (3) are considered a marked rheme since the finite verb *came* exceptionally appears before the grammatical subject in the declarative sentence. This highlights the aforementioned divergent word order within the rheme that merits marked rheme status. Sentence (3) is also reminiscent of the Presentation Scale by Firbas, whereby *a strange man* would act as a rheme due to its PHENOMENON PRESENTED status in the scale. Hence, similar statements that employ presentational or appearance-based verbs (e.g.,

[10] As a reminder, Weil and Mathesius also employed the concept of markedness, albeit with the term *pathetic*. From Firbas and beyond, the term *marked* was employed.

come, appear, occur, turn up, happen, arrive, come up; see Chapter 2.3) would force end-shifted subjects and thereby marked rhemes.

In (4), the nominal predicate *a number of interesting points* is shifted to the end of the rheme. The prepositional phrase *in this article* is inserted between the thematic existential *there were* and the nominal predicate *a number of interesting points*. As the standard position of circumstantials is at the beginning or end of the sentence, the prepositional phrase *in this article* would be considered the marked rheme and *a number of interesting points* the unmarked rheme (Taglicht 1984: 23–24).

A similar phenomenon occurs in instances such as (5) where insertion of an element, typically an adverbial or adjunct, interrupts the structure of the rheme to cause it to be marked. In (5), *however* could have appeared sentence initially or finally. The former case would make it a marked textual theme; realizing *however* at the end of the sentence would afford it rhematic status. Yet, its medial realization between the finite verb and complement is considered divergent rhematic word order. As such, the cohesive tie *however* becomes a marked rheme while the remaining sentence constituents remain unmarked.

As for Dubois, she argues for the identification of marked rhemes with the following stipulation: The rhematic sentence constituent realized to the left of its unmarked position in the sentence becomes marked; conversely, marked rhemes are sentence constituents realized sentence-finally if no insertion is present (Dubois 1987: 106). It is the second condition that accounts for the marked rheme in (3) above, such that *a strange man* is considered the marked rheme. Additionally, *a number of interesting points* in (4) and *to England* in (5) are the unmarked rhematic positions due to their respective insertions *in this article* and *however*, respectively. Since these insertions appear to the left of the unmarked rhemes, they become marked rhemes. Dubois adds that the communicatively most dynamic element, in terms of informational content, should also be placed to the left of finite verbs (the unmarked rheme position) for marked rheme status. The transposition of the predicative elements to immediately after the finite verb accentuates their dynamicism and marked rheme status (Dubois 1987: 106).

A distinction between marked and unmarked rhemes remains absent from Halliday's framework as all rhematic constituents are considered to equally contribute to the development of the discourse message. Whereas marked themes emphasize their discursive effect on text structure and development (e.g., contextualization of discourse statements with circumstantial adjuncts *in past research* or *under present conditions*), rhemes are analyzed in terms of how the NEW informational and propositional content brings discourse forward.

Halliday's treatment of NEW and GIVEN information vis-à-vis thematic and rhematic constituents represents a further divergence from the Prague School. Prague School linguists, such as Mathesius (1983), van Dijk (1977), Sgall et al. (1973), Daneš (1974) and Firbas (1992), generally conflate the theme with GIVEN and the rheme with NEW;[11] Halliday, however, separates the two systems: "The theme is what I, the speaker, chose to take as my point of departure. The Given is what you, the listener, already know about or have accessible to you" (Halliday & Matthiessen 2014: 116). He further argues that NEW elements are first and foremost those which have already been mentioned explicitly in the text previously. Language users can additionally afford exceptional, 'newsworthy' status to a sentence constituent for it to be interpreted as NEW. While this is much more readily accomplished in spoken speech through

[11] Note that Prague School linguists allow NEW themes, such as at the beginning of narratives; however, upon determining the degree of communicative dynamism (CD), the theme will ultimately have the lowest CD despite its NEW, i.e., context-independent, status (cf. Davidse 1987: 65).

the tonic foot (see Halliday & Matthiessen 2014: 116–118 for a discussion of establishing NEW elements in spoken speech), clear-cut markers in written speech are far less overt.

Halliday's approach alludes to the necessity of NEW but the optionality of GIVEN elements in text. If GIVEN elements are indeed present, such elements are said to be phoric, which is comparable to Firbas' distinction between context-dependent (GIVEN) and context-independent (NEW) information (see Chapter 2.3). By and large, marked thematic elements typically indicate a NEW element as the marked status inherently fulfills an emphatic function. Exceptional cases, such as existentials, clefts, thematic equatives and fronted elements, also default to NEW information due to their marked status. Otherwise, upholding or deviating from standard word order alone is the only way to consistently identify what should be identified as NEW in written speech. Tracing the development of discourse via GIVEN and NEW discourse topics is thus embodied in the realization of themes and rhemes throughout a text. This interplay between theme, rheme, GIVEN and NEW reveals the complexity behind information status, propositional content and discourse development as formalized in Halliday's conceptualization of thematic theory.

Halliday has remained one of the most prevalent contributors to contemporary research on the thematic paradigm ever since his initial work in 1967. Despite his divergent approach from that of the Prague School, his work continued to be informed by its adherents. Daneš and Halliday in particular show significant inspiration from each other's work on thematic progression and a continued development of the thematic paradigm. Through his work with Hasan on the topic of cohesion, Halliday formalized fundamental changes to how structural, textual and discourse analysis is conducted in the field of text linguistics. Considerable contemporary work on the thematic paradigm within the Hallidayan framework indicates the far-reaching implications his conceptualization has had on a modern understanding of thematic structure from the perspective of systemic functional grammar.

2.6. Excursus: A Brief Survey of Historically Problematic Theme Boundaries

Building off the normative approach to theme/rheme identification from the previous sections, the present section addresses problem cases in thematic theory from both a Prague School and Hallidayan perspective. The cases in question are existentials, predicated themes and clefts, as illustrated in the respective example sentences in Table 2-2 taken from the Corpus of Contemporary American English (Davies 2008-).

	THEME PROPER	DIATHEME	TRANSITION	RHEME	RHEME PROPER
Prague School (FSP)	LOWEST COMMUNICATIVE DYNAMISM →			HIGHEST COMMUNICATIVE DYNAMISM	
	(1) There		*were*	*some no-go zones*	
	(2) It		*is*	*women*	*who disproportionately [...]*
	(3) It		*is*	*clear*	*that the [...]*

Table 2-2: Thematic constituent analysis of existential (1), predicated themes (2) and clefts (3) according to the functional sentence perspective approach and with the help of communicative dynamism.

Starting with the Prague School approach by determining the communicative dynamism of the sentence constituents, the principle of linearity applies since standard word order is present; hence, an increasing degree of communicative dynamism emerges while progressing from the beginning to the end of each text. In (1) specifically, the verbal transition element *were* has a lower communicative dynamism than the context-independent object *some no-go zones*,

thereby attributing the object rhematic status. Additionally, *there were* belong to the PRESENTATION OF PHENOMENON while *some no-go zones* equates to the PHENOMENON PRESENTED (see Chapter 2.3). In the Presentation Scale, this is an exact reflection of the scale's interpretive arrangement and coinciding dynamic semantic function. Hence, *there* has the lowest communicative dynamism and functions as the theme in the sentence. While not outlined formally here, the Combined Scale is then used to determine the thematic and rhematic constituents of (2) and (3) based on their communicative dynamism. This results in the thematicity detailed for predicated themes and clefts in Table 2-2.

In the Hallidayan approach, since *there* is neither circumstantial nor participant, it takes on the exceptional interpersonal subject status due to its process function of indicating existence (Halliday & Matthiessen 2014: 308). The existential *there* alone constitutes the (topical) theme and the remainder becomes the rheme (cf. Table 2-3). The same can be said about the predicated theme and cleft in (2) and (3), respectively. There, the dummy-*it* has interpersonal subject status and is given thematic status from a Hallidayan perspective.

Sentential Bifurcation	α-THEME		β-RHEME	
Existential	*(1) There*		*were some no-go zones.*	
Clausal Bifurcation	α-THEME	α-RHEME	β-THEME	β-RHEME
Predicated Theme	*(2) It*	*is women*	*who*	*disproportionately carry the burden of its absence.*
Cleft	*(3) It*	*is clear*	*that*	*the generations interact with each other.*

Table 2-3: Hallidayan division of thematic and rhematic structures in α and β-clauses of existentials, predicated themes and clefts as derived from Halliday & Matthiessen 2014: 124.

Within the Hallidayan framework, clefts and predicated themes are afforded an additional level of abstraction due to the requisite β-clause that accompanies the matrix α-clause. For clarification, a β-clause corresponds to the subordinate clause and an α-clause corresponds to the independent clause of a sentence. Further abstraction splits each clause into their own thematic and rhematic parts, instead of one set of themes and rhemes for the entire sentence, as shown in Table 2-3.

Through the two sets of themes and rhemes, the standard bifurcation of thematic and rhematic elements is highlighted: the dummy-*it* forms the α-theme and the predicate becomes the α-rheme; the relative pronoun *who* and subordinating *that* in the β-clause then become the β-theme, with the remaining sentence constituents of the β-clause forming the β-rheme. The exception to thematicity appears at the sentence-level bifurcation when the entire α-clause becomes the theme and the entire β-clause forms the rheme.

The rationale for exceptionally extending the theme to include the predicate of the α-clause can be justified with the topical theme at the end of the α-clause (i.e., *it is women*), as is the case in (2). The lines begin to blur somewhat in (3), since the first topical theme does not emerge until *the generations* in the β-clause. The solution to this issue with predicated themes, clefts, and existentials by extension, is the function that these structures fulfill discursively. In clefts and predicated themes, the dummy-*it* is neither deictic nor coreferential. It does not function as an anaphoric reference to a constituent in the previous sentence nor to referents within the discourse. Similarly, the existential *there* is not deictic either since it does not spatially or temporally refer to anything within the discourse.

Since the existential *there*, clefts and predicated themes cannot rely on a denotative foundation, on which the remainder of the message builds (see Breivik 1981, Langacker 1987), their discursive function must account for their thematic status instead. Through their discursively cataphoric nature, the existential *there* and dummy-*it* both make the rhematic elements most salient. They indicate to text recipients that they should pay particular attention to the informational content in the rheme (Quirk et al. 1985: 1402). Discursively, this builds tension upon instantiation of either the existential *there* or the cleft's dummy-*it*, which is then resolved once the rheme has been realized (see also Collins 2015). The β-clause rheme *who disproportionally carry the burden of its absence* in (2) answers the incomplete proposition put forward in the message of the theme *It is women*. Similarly, the β-clause rheme *that the generations interact with each other* in (3) resolves the informational tension established through the thematic proposition in the α -clause *It is clear*.

Regardless of a bipartite or multipartite analysis, the dummy-*it* in clefts and predicated themes first and foremost functions discursively as an attention marker to the reader. The α-clause establishes tension to be resolved in the β-clause, whose theme returns to a context-dependent, i.e., retrievable, foundation. The β-clause rheme ultimately builds on the thematic foundation within the β-clause to move the message forward. The benefit to a multipartite analysis is greater analytical detail: the contribution that each theme and rheme pair affords to the communicative message can be readily traced. However, by splitting theme-rheme pairs into additional theme-rheme pairs specifically for clefts and predicated themes (but not for existentials), the discursive function that these structures afford becomes lost. Particularly because these structures lack a coreferential function and therefore fail to fulfill the requirement of the theme as the propositional foundation of a statement, a bipartite, sentential approach would accommodate their exceptional structure and discursive function. Employing a sentential bifurcation of thematicity, as shown in Table 2-3, allows the discursive function of these structures to be encoded in their syntactic structure.

This approach is advocated by Thompson (2004), who argues:

> […] the function of a predicated theme is to single out the predicated constituent as particularly noteworthy in some way, often because it contrasts with something in another part of the text […], or because it is represented as selected from amongst a number of alternatives." (Thompson 2004: 156)

This is similar to how Leong (2005) treats clefts, thematic predicates and existentials, whose thematic analysis is exemplified in (4). By extending the thematic span to the entirety of the existential phrase *there is*, the existential's discursive function is mapped onto both the existential construction and the theme as a single unit.

(4) **There is** little evidence for such unfounded claims.

Davies (1997: 61) takes this approach even further by including the existential construction *there* + *copula* + *experiential subject*, as in (5).

(5) **There is little evidence** for such unfounded claims.

While functionally logical, as the writer employs the existential *there* to introduce a NEW theme through the experiential subject, this analytical approach would result in sentences such as *There is no time* possessing no rheme. This could pose an intriguing rhematic break, in contrast to the well-established thematic break: it could serve a similar function of beginning a new

discourse topic as a hypertheme or could derive a NEW topic from a previously established one. However, it has become commonly accepted that sentences must contain a rheme at the bare minimum even if no theme is realized. Therefore, most researchers fall either into the Hallidayan camp of a sentence-level bifurcation as outlined in Table 2-3 or an exceptional treatment as put forward by Thompson (2004).

A final note on thematicity with existentials, clefts and predicated themes is that their exceptional treatment can be circumvented entirely with the communicative dynamism model, as shown in Table 2-2. Instead of relying on metafunctional topical themes, the principle of linearity, dynamic semantic function and contextual constrains can determine the thematicity of each sentence constituent. That being said, this cannot inherently account for the dummy-*it* or existential *there* failing to serve as a communicative foundation for subsequent rhematic information; the discursive function of these structures as realized thematically therefore remains obscure in a Prague School interpretation.

As will be detailed in Chapter 4, the present research follows the approach forwarded by Thompson (2004) in order to map the discursive function of these exceptional structures onto their thematic realization patterns. Their exceptional status within thematic theory is therefore given exceptional treatment in their analysis as formalized in the present work.

2.7. Summary of Theme and Rheme per the Prague School and Hallidayan Framework

The present section summarizes the most important tenets of the thematic paradigm as derived from the research presented thus far. Here, both the Prague School of Linguistics and the Hallidayan approach to thematicity and its determination are presented.

	In the coming years,	*Mr. Grave*	*will clearly*	*have*	*greater responsibility*	*for others*
Prague School (FSP)	**THEME**		**RHEME**			
	THEME PROPER	DIATHEME	TRANSITION PROPER	TRANSITION	RHEME	RHEME PROPER
	SETTING	BEARER	QUALITY		SPECIFICATION	FURTHER SPECIFICATION
	LOWEST COMMUNICATIVE DYNAMISM →				HIGHEST COMMUNICATIVE DYNAMISM	
Hallidayan Approach (SFG)	MARKED	UNMARKED				
	TOPICAL THEME	RHEME				

Table 2-4: Summary of thematic and rhematic constituents in textual realization according to the two schools of thought, functional sentence perspective (FSP) and systemic functional grammar (SFG).

Within the Prague School framework, communicative dynamism determines which sentence constituents belong to the theme, transition or rheme. Communicative dynamism is determined by a combination of the principle of linearity, dynamic semantic function and context. The principle of linearity states that sentential elements increase in communicative dynamism when progressing from the beginning to the end of the sentence. This only applies when a sentence follows standard word order. Otherwise, the dynamic semantic function is used to determine communicative dynamism with the help of the Presentation, Quality or Combined Scale. Finally, communicative dynamism is motivated by context. Context-dependent elements, i.e., elements

immediately retrievable from the situational context or cotext, are ascribed lower communicative dynamism and are nearly always thematic (e.g., *In the coming years, Mr. Grave* from Table 2-4). Sentence constituents with the highest communicative dynamism are rhematic (e.g., *greater responsibility for others* from Table 2-4). Those that fall in between the upper and lower bounds of a sentence's communicative dynamism are then the transition (e.g., *will clearly have* from Table 2-4). In terms of information status, thematic elements are nearly always conflated with GIVEN information and rhematic elements with NEW information. The conflation of thematicity and informational status according to the Prague School is known as the combining approach.

When multiple thematic, transition and rhematic elements are present, they can be broken down even further into theme proper, diatheme, transition proper, transition, rheme and rheme proper (cf. Table 2-4). The thematic element with greater communicative dynamism qualifies as the diatheme, the element with lesser communicative dynamism qualifies as the theme proper. The boundary between the theme and rheme is given by both the transition and the transition proper. The latter carries greater communicative dynamism than the transition; both elements correspond to the verb(s) of the sentence, either notional or categorial. After the transition comes the rheme proper, typically the final element in a sentence and that with the highest communicative dynamism. The sentence constituent whose communicative dynamism is higher than the transition but lower than the rheme proper is then the rheme

Finally, despite markedness having originally stemmed from the Prague School, the term *marked theme* is rarely used. Instead, a theme's markedness is expressed through the bipartition of the theme proper and diatheme. The theme proper is typically considered the marked thematic element, whereas the diatheme is unmarked. The diatheme can be further subdivided into diatheme and diatheme-oriented as well but is not universally advocated (and therefore not reflected in Table 2-4). In such instances, the former is marked, the latter unmarked. Ultimately, however, it is recommended to consider markedness in terms of degree of communicative dynamism within the Prague School framework.

While communicative dynamism underlies the thematic paradigm in functional sentence perspective, metafunctions are central to the Hallidayan approach. These can be broken down into textual, interpersonal and experiential metafunctions, of which the first two are optional. Textual themes aid in establishing cohesion and coherence across sentences. Interpersonal themes embed modality and mood within the sentence to express the speaker's stance toward the message's propositional content. Topical themes are typically either participants if they constitute the grammatical subject of a sentence or they are circumstantial adjuncts to the verb.

Regardless of the number and kind of themes realized through the metafunctions, every sentence must contain a topical theme, which is the first experiential topic of the sentence. These are typically circumstantial adjuncts (*In the coming years* in Table 2-4) or the grammatical subject. When experiential topics are realized before the grammatical theme to become the topical theme, they constitute the marked themes of the sentence. The grammatical subject then becomes relegated to the rheme (*Mr. Grave* in Table 2-4). If the grammatical subject is the first experiential topic, then it becomes the unmarked topical theme. Everything after the topical theme then qualifies as the rheme.

Finally, systemic functional grammar adherents do not conflate GIVEN information status with the theme and NEW information status with the rheme. Instead, thematic elements may explicitly be ascribed NEW status, such as in the first sentence of a text. A GIVEN rheme in written speech can be realized through the exact lexical repetition of previously established discourse topics

across concomitant rhemes. Otherwise, phoric elements realized rhematically maybe also be attributed GIVEN status due to their context dependence through previous explicit realization within the discourse. In both spoken and written speech, fronted elements can be used for emphatic or corrective purposes to afford typically GIVEN themes NEW information status. Marked themes inherently have NEW information status in their realization, particularly when belonging to the interpersonal and certain textual metafunctions.

The greater analytical complexity in the determination of information status and thematicity with the Prague School approach may allow for a closer delineation of the communicative weight that each sentence constituent affords. However, even with the more simplistic Hallidayan approach, the distribution of thematic and rhematic elements vis-à-vis information status and contribution to discourse can be systematically isolated. Upon determination of the thematic constituents and their progression, the discursive function they fulfill as a textual and structural tool in the unfolding of discourse can be scrutinized.

Chapter 3 – Theme as a Textual and Structural Tool

Chapter 2 first and foremost outlined the development of the thematic paradigm and the various ways themes and rhemes can be identified in a text. The function, purpose and conscious development of the theme, however, has only received secondary treatment in the present work so far. The purpose of Chapter 3, therefore, is to scrutinize the structural, textual, discursive and systematic function of the theme as a tool for communicative development in written text.

Specifically, thematic selection as a factor of thematic progression and the so-called method of development forms the initial basis of the discussions presented in this chapter. The relationship between thematic selection, thematic progression and discourse development is then explored as a potential characteristic of textuality and text type. Finally, computational approaches to thematic theory and automated thematic constituent analysis as forwarded in contemporary research are outlined.

3.1 Thematic Selection and Progression

At this point, it may be worth recapitulating what both the theme and rheme contribute to the development of communication. The theme has come to be defined as:

- The point of departure, but one of deliberate and conscious choice (Weil 1978, Daneš 1974)
- The foundation or starting point of the utterance (Mathesius 1983, Thompson 2004)
- The element which contributes least towards the development of communication (Firbas 1996)
- The peg on which the message is hung (Halliday 1970)

The rheme has come to be defined as:

- The goal of discourse (Weil 1978, Mathesius 1983)
- The element which contributes greatest towards the development of communication (Firbas 1996)
- The part of the clause in which the theme is developed (Eggins 2004, Halliday & Matthiessen 2014)

The theme is therefore the (contextually) established foundation of a message the language user wishes to convey, the communicative bedrock upon which a writer sets the stage. With deliberate selection of a theme, a speaker specifies the initial direction in which a message will unfold. Where the message ultimately lands, how it comes into communicative fruition and unfolds as purposeful communicative development is the function of the rheme. The rheme is the answer to the informational question posed by the theme: Where does the writer intend to take this theme? What delimiting (rhematic) information does the speaker provide on the basis of how the text started? The rheme thus brings the communication forward by manifesting a clause's discourse goal and by contributing to the message's development. Considering a discourse unit on the whole, there is an interplay between the message's foundational theme and expositional rheme as the text progresses. Such continuous interplay between a text's thematic and rhematic elements becomes salient through its thematic progression (Daneš 1974) and so-called method of development (Fries 1995).

3.1.1 Thematic Progression: From Daneš to Modern Interpretations

Daneš (1974) defines thematic progression as a reflection of the overall thematic structure of a text. How an author successively progresses the text with each clause is reflected in deliberate thematic selection. Tracing the thematic selection throughout a text thus reveals its thematic progression. In other words, the choice of theme from clause to clause (consciously or otherwise) is influenced by previous instantiations of thematic or rhematic elements. This final collective of individual choices of how to begin and move a communicative message forward is therefore embodied in thematic progression. It is important to note that thematic progression as a system was not formalized until Daneš's formal treatment of the topic. Before then, identification and treatment of thematic and rhematic elements within a single clause was the focal point of research, less so how and why these individual elements were then *developed* throughout a discourse unit. Daneš's greatest contribution was thus his formalization of thematic progression as a means to tracking the overall structure of a text and its communicative development.

In his seminal work *Functional sentence perspective and the organization of the text,* Daneš (1974) outlined four fundamental forms of thematic progression: 1) **simple linear progression**; 2) **constant (or continuous) theme**; 3) themes derived from a **hypertheme**; and 4) **split rheme progression**. Examples of these progression patterns will be treated in the following and were taken from the Corpus of Contemporary American English (Davies 2008-).

The first form of thematic progression is **simple linear progression** whose partial themes (T_n) and rhemes (R_n) that instantiate progression are given in bold:

(1) Granger causality is a **notion** (R_1) based on the ability to predict the future value of one process using the past values of another process. This **notion** ($R_1 \rightarrow T_2$) was first introduced in macroeconomics and has proven useful in providing the direction of information flow.

(2) And inflation would have been even weaker but for the impact of **Mr. Trump's** (R_1) tariffs, according to a study published Friday by the Federal Reserve Bank of New York. **Mr. Trump** ($R_1 \rightarrow T_2$) has repeatedly argued that the cost of the tariffs, which are taxes on imported goods, is being absorbed by Chinese exporters.

Simple linear progression is characterized by the development of a rhematic element from one sentence and its realization as the theme in the subsequent sentence, denoted ($R_1 \rightarrow T_2$). In (1), *notion* first appears rhematically as a means to introduce NEW information to the foundational, thematic topic of *granger causality*. In (2), *Mr. Trump's* functions as the rheme in the possessive in the first sentence and is taken up as the thematic grammatical subject in the second sentence. In other words, NEW information is introduced as the rheme of the first sentence and realized thematically to establish the foundation of the subsequent sentence's message. Instantiation of previously established topics has no upper limit but is generally constrained to fall between three and seven sentences prior (McCabe 1999: 176, Svboda 1991: 88–89). While both (1) and (2) employ lexical repetition as cohesive devices to instantiate thematic progression, coreference relationships such as synonymy, hypernymy, ellipsis as well as lexical entailment are also a common means of thematically progressing communication from one sentence to the next.

It should be stressed that any rhematic element from the immediately preceding sentence can be employed as the theme in the subsequent sentence with simple linear progression (or any thematic progression pattern, for that matter). The reasoning for this again lies in the nature of

the rheme: Its purpose is to constitute the core of the discourse message which is based on the thematic foundation of the message. In doing so, it develops the communication at hand. Upon initial rhematic realization, any mention of the same rhematic elements within the subsequent theme provides an informational link to previously established topics. This functions similar to lexical or cohesive chains (cf. Halliday & Hasan 1976), whereby the reader can trace instantiations of the same discourse topic, i.e., theme or rheme, across sentences. In simple linear progression, a back-and-forth or zig-zag pattern emerges that structurally and textually expresses the dynamicism behind the informational content and its progression (Rosa 2013: 221).

The next form of thematic progression is that of **constant (continuous) theme progression**. Here, the theme from one sentence is re-instantiated as the theme of the immediately following sentence. In both (3) and (4), the same theme is realized through lexical repetition in nearly every sentence. The sole exception is *this* (or even *game*) being referred to in the second clause through the proform *it*. Similar to the example with simple linear progression, constant theme progression can be instantiated through lexical repetition, lexical entailment (i.e., paraphrase, synonymy, hypernymy) or coreference.

(3) **This** (T_1) is a game, **it's** ($T_1 \rightarrow T_2$) about fun. But for **it** ($T_2 \rightarrow T_3$) to work, **it** ($T_3 \rightarrow T_4$) has to be shared fun.
(4) In this report, the term '**impacts**' (T_1) is used primarily to refer to the effects on natural and human systems of extreme weather and climate events and climate change. **Impacts** ($T_1 \rightarrow T_2$) generally refer to effects on lives, livelihoods, health, ecosystems, economies […].

Instead of expounding on informational content from a previous rheme, a previously established theme is given further elaboration. Repeated use of the same theme thereby allows tracing the foundation of the message in a text more readily as opposed to its discursive development found in the rheme. This type of thematic progression is typically found in narratives or descriptions, and may run the risk of the text sounding dull, repetitive or predictable (Rosa 2013; Rørvik 2012; Hawes 2001). That being said, constant theme progression allows rigid and structurally sound texts to emerge since the reader and writer can consistently make use of and trace the same theme.

(5) [1] For many **economists** (H_1), the current wave of automation anxiety amounts to misguided scaremongering by modern-day **Luddites** (H_2). [2] After all, they point out, the prediction that automation will supplant **human labor** (H_1) on a massive scale has recurred in both Utopian and dystopian flavors throughout the history of industrialization. [3] **Futurists of the past** (H_2) have predicted that mass automation will usher in an era of human liberation from toil, or that it will immiserate all but the fortunate few who own or create the machines. [4] Time and again, however, the **economy** (H_1) has defied such predictions. [5] For centuries, automation has been destroying some **jobs** (H_1) while creating other jobs -- usually better paid and less grueling -- and driving economic growth and prosperity. [6] In short, the history of automation's impact on the **labor market** (H_1) has been one of "creative destruction," a mantra to which many economists adhere today.

The third form of thematic progression Daneš established builds upon Firbas' initial conceptualization of the so-called **hypertheme**, from which subsequent themes are derived. In (5), all sentences in the passage are linked by the hypertheme (H_n) *economy/economics*, from which *automation* is derived and repeated. The only sentence that proves an exception to this

type of thematic progression is sentence [4], which forms what Dubois came to call simple gapped derived from hypertheme (cf. Dubois 1987). Here, the hypertheme *economy/economics*, not the repeated theme *automation*, is realized. Additionally, the second hypertheme *Luddites* is established in the second sentence and contextualized around the theme *automation* and the hypertheme *economy/economics*. These two hyperthemes form the overarching topics through the paragraph, although their discursive treatment is not limited to the paragraph level alone.

Hyperthemes thus establish a form of scaffolding, whereby associated, often hypernymous, terms are instantiated thematically. The hypertheme *economy/economics*, in fact, is realized in every sentence through varying word forms, hypernymy or collocational relationships, e.g., *economist, mass automation, human labor*, *jobs* and *labor market*. It should also be noted that the text in (5) otherwise reflects a nearly consistent constant progression. This is achieved through *automation* appearing as the theme in the independent and dependent clauses in most sentences. Thematic progression ensuing conjunctively, here through constant progression and hypertheme, shows how the hypertheme affords a greater macro structure, with constant progression forming the microstructure of a text. The interplay between the hyperthemes and themes reinforces not only the structural framework of this text but also the development of its informational content.

This form of thematic progression has been subject to considerable criticism given its close relation to the concept of discourse topic, albeit without consideration to how the topic is textually developed. Further, some linguists argue for a more simplistic approach through thematic analysis as either simple linear or constant theme progression under the assumption of a shared discourse topic (cf. Witte 1983; Dubois 1987; Leong 2005, 2007).

In Daneš's final thematic progression pattern, **split rheme progression**, multiple discourse topics are presented as enumerated or coordinated rhemes and then instantiated as individual themes in subsequent clauses (cf. (6) below). The first rheme in the list, *elementary substances,* becomes the theme of the second sentence (R_{1a}→T_2). The second rheme in the list, *compounds*, in turn, becomes the theme of the third sentence (R_{1b}→T_3). Hence, the original rheme is split and treated thematically as the text progresses. This kind of thematic progression is most common in academic, scientific and formal writing, where concepts are initially introduced to the reader and then explained individually (Jalilifar 2009: 98). Due to the proximity between rheme and themes that subsequently follow, the reader is able to retain the rhemes in short-term memory and access them as GIVEN information later in the text.

(6) All substances can be divided into two classes: **elementary substances** (R_{1a}) and **compounds** (R_{1b}). An **elementary substance** (R_{1a}→T_2) is a substance which consists of atoms of only one kind […] A **compound** (R_{1b}→T_3) is a substance which consists of atoms of two or more different kinds […] (Daneš 1976: 121)

What is noteworthy in split rheme progression is the microstructure it invariably produces. Since split rheme progression introduces multiple topics that are then elaborated upon in subsequent themes, a combined structure reminiscent of hyperthemes and simple linear progression emerges: NEW discourse topics derived from the defining discourse topic of the paragraph, i.e., hypertheme, are first introduced in the rheme of the sentence. Then, each rhematic topic is treated individually throughout the paragraph. By linearly instantiating the rheme as subsequent themes, the author is free to vary the brevity or depth of the remaining text concerning said theme. If a single clause for each respective rheme follows, as is the case in Daneš's example from (6) above, then split rheme progression is present. Should the author choose to employ multiple clauses for each respective rheme, then additional progression

patterns within the split rheme progression emerge. This latter case results in a more complex microstructure within the paragraph itself.

To illustrate how split rheme progression can introduce this microstructure and to make the previous explanation more concrete, the following excerpt may prove helpful:

(7) Brand awareness can be split into two categories, **family branding** (R_{1a}), and **individual branding** (R_{1b}). **Family branding** (R_{1a}→T_2) is where several different products are marketed together, under one larger umbrella brand. Examples of **family branding** (T_2→T_3) include Ford Motor Company, Apple Computers, and Bath and Body Works. **Individual branding** (R_{1b}→T_4) is where each product is promoted as its own separate entity, such as Absolute Vodka, Eggo Waffles, and Oreo Cookies. When choosing a **marketing strategy** (H_1) that is right for your company it is best to make an informed decision by exploring the different upsides and drawbacks to each strategy. (Pritchard 2012)

In the initial sentence, two rhematic elements are introduced topically as categories: *family branding* and *individual branding*. This alludes to, but does not necessitate, split rheme progression since both rhemes must be realized in subsequent clauses later for split rheme progression. The first rhematic category *family branding* is realized as the theme in the second sentence through simple linear progression (R_{1a}→T_2). The same theme *family branding* is developed in the third sentence through continuous progression (T_2→T_3). Treatment of the topic *family branding* through the patterning R_{1a}→ T_2→ T_3 encapsulates the first microstructure that split rheme progression instantiates. In the second microstructure, the second rheme *individual branding* becomes the theme of the fourth sentence (R_{1b}→T_4). It is at this point that split rheme progression finally manifests itself. The paragraph concludes with the hypertheme *marketing strategy* as the theme since it is a hypernym of *individual branding* and *brand awareness*. The use of the paragraph's hypertheme in the final sentence affords a summarizing or concluding function to the treatment of the discourse topic *brand awareness* in the text.

This paragraph exemplifies the variability in clause distance between initial introduction of listed rhematic elements and their later instantiation as themes through split rheme progression. While Daneš's original model treated the two rhemes in a single sentence each, authors can produce multiple sentences for each rheme before progressing onto the next rheme. That way, split rheme progression in particular can more clearly indicate the beginning of new microstructures within a text once enumerated rhematic elements are thematically realized later.

Daneš's original conceptualization of thematic progression patterns has since been expanded to account for unexplored patterns and identified deficiencies. The first addition to Daneš's thematic progression models is advocated by Dubois (1987), McCabe (1999) and Li (2009) and is called **split theme** or **multiple-theme progression**. The pattern is reminiscent of the split rheme pattern but revolves around themes alone. It thus functions as the thematic counterpart to split rheme progression by introducing multiple thematic elements that are later realized as individual themes. This can be seen in (8a) – (8c) as illustrated in McCabe (1999: 175):

(8a) The upward movement of **wages** (T_{1a}) and the downward **price of cereals** (T_{1b}) led […]
(8b) Better **wages** (T_{1a}→ T_3) in both town and countryside enabled the population to […]
(8c) While the **price of wheat** (T_{1b}→ T_4) fell, wine, beer, oil, butter, cheese, meat, fruit, […]

Here, the two themes *wages* and *price (of cereals)* are introduced in (8a) and realized individually as themes in (8b) and (8c). Microstructures within split theme progression can similarly emerge and are akin to the microstructuring afforded by split rheme progression. With this progression pattern, the foundation, not the core, of the message is given priority through repeated realization of previously enumerated themes.

The second addition to Daneš's progression patterns stems from split theme/rheme progression and the gaps that often emerge within this progression type. If thematic or rhematic elements are instantiated two or more sentences after their initial introduction, then the progression pattern becomes that of **simple gapped progression** (Dubois 1987, Hawes 2015) or **complex linear/continuous progression** (Rørvik 2012). In contrast to split theme/rheme, neither enumerated nor multiple theme/rhemes are required in order to make use of simple gapped progression/complex linear progression. Rørvik exemplifies this gapped pattern in (9a) – (9c), whose themes are in bold and gapped themes underlined.

(9a) In a way **I** (T_1) feel sorry for people with no <u>dreams</u> (R_1).
(9b) **I** ($T_1 \rightarrow T_2$) think they are missing out on something for what could possibly feel better than accomplishing your goals?
(9c) <u>Dreams</u> ($R_1 \rightarrow T_3$) are forever, whether big or small!

Constant progression exists between sentences (9a) and (9b) through *I* being instantiated thematically in both. However, the theme *dreams* in (9c) stems from the earlier rheme in (9a) with a gap of two sentences. The discourse message around *dreams* therefore remains undeveloped until sentence (9c). In between, the intermediary sentence (9b) is inserted as an anecdotal, secondary treatment of the topic introduced in the first sentence.

There are two important points concerning the employment of simple gapped progression. Firstly, it is not limited to anecdotal insertions alone. Particularly in scientific text, Dubois claims that to further elucidate the theme or rheme initially introduced, "investigators are dealing with a complex set of interrelated ideas, not with a simple, straight-line narrative, and […] the complexity of thematic development necessarily reflects that of scientific content" (1987: 95). Secondly, gapped progression patterns may indeed have multiple intermediary clauses. While no upper bound has been posited, one naturally runs the risk of overcomplicating a text's structure if the reader must return to syntactically distant clauses to identify a theme's referent.

In order to account for this, researchers have further refined what can constitute gapped progression, with differing terminology. Berry qualifies wide-gapped progression as **discourse themes** (1995: 18), McCabe as **peripheral themes** (1999: 180–181), Rørvik as **complex continuous progression** (2012: 168) or **extended reference** (2012: 169), Hawes as **constant gap** (2001), and Jingxia & Li as **summarized progression** (2013: 120). Where gapped progression is evident, it is important to note that thematic progression does not simply cease in the intermediary sentences. Similar to how a microstructure manifests within split rheme progression, the sentences reflecting gapped progression form the beginning and end boundaries to the ensuing microstructure of the intermediary sentences.

The final extension to Daneš's thematic progression types is that of **constant *type* progression** from Hawes (2010b). This functions similar to constant continuous progression, in that the theme from one clause is reiterated as the theme in the subsequent clause. The primary difference between the two is that the themes follow the same grammatical types, such as reiterated existentials, clefts, prepositional phrases or WH-question words (Hawes 2010b: 47).

The last case of reinstantiated WH-question words as themes is shown in (10). There, *what* and *how* form the thematic foundation of the discourse message and thereby establish structural parallelism in how both sentences are syntactically realized.

(10) **What** (T_1) a Christmas present it would be. And **how** ($T_1 \rightarrow T_2$) [Gordon Brown]'d love to play Santa! (Hawes 2010b: 47)

Constant type progression can be used for emphatic purposes, as is the case in (10), but is typically employed to increase clarity, draw attention to the topic and establish a rhythm in both sentences (Kazemian & Hashemi 2014: 1183). Due to the resulting structure, constant type progression is akin to structural parallelism across sentences. This establishes regularity and predictability in the presentation of thematic discourse topics, which eases comprehension of the content being developed (Kazemian & Hashemi 2014: 1183).

The previous additions to thematic progression patterns are ultimately further refinements of Daneš's original models. The reason for their (continued) conceptualization has been to account for the greater complexity that language users can afford to the structure and development of their text. Contemporary models can more readily reveal how topics are cogently developed and illuminate the underlying structural and rhetorical patterns of a text. These may have otherwise been neglected or remained less overt with Daneš's models alone. Both the original thematic models and the contemporary additions are now considered standard inventory of thematic progression patterns. Still, there remain certain models that do not belong to this inventory and concern themselves with the instantiation of elements in the rheme only. As these patterns account for the less considered rhematic progression and have been adopted in the present work, they will be addressed briefly in the following section.

3.1.2. Less Explored Progression Types: Rhematic Development

When reflecting on the patterns outlined in the previous section, the absence of two progression models is noteworthy: progression of the theme in the first clause to the rheme in the subsequent clause ($T_1 \rightarrow R_2$); or progression of the rheme in the first clause to the rheme in the subsequent clause ($R_1 \rightarrow R_2$). Advocates of these models are Enkvist (1973), Cloran (1995), Shi (2013) and Dou & Zhao (2018).

(1a) For a time, **they** (=beavers) (T_1) were in danger of disappearing completely.
(1b) But laws were passed to protect the **beaver** ($T_1 \rightarrow R_2$).

Rhematic regression,[12] as exemplified in (1), illustrates how *they* transitions from the theme in (1a) to the rheme in (1b) as *beaver* (Shi 2013: 1641). A $T_1 \rightarrow R_2$ progression indicates that the foundation of a message is subsequently developed as the core of a message, which may prove conceptually controversial. Three possible explanations for this type of progression can be provided through the GIVEN/NEW paradigm, through so-called peripheral themes, and through rhetorical functions in writing.

Starting with GIVEN/NEW, Prince (1981) and Kopple (1991) outline additional factors to account for the information structure in rhematic regression specifically. Firstly, the theme-GIVEN, rheme-NEW assumption is switched. In the second sentence (1b) above, the theme *laws* does not have GIVEN but rather NEW status; at the same time, the rheme *beaver* has GIVEN status.

[12] Note that the development of elements into rhemes across sentences is denoted as rhematic regression, not progression. To establish a rhematic parallel to thematic progression, however, the present work employs the term rhematic progression in subsequent sections.

In order to highlight this marked information structure, Prince and Kopple postulate three classes of NEW information status: NEW, BRAND-NEW and UNUSED. Of these, BRAND-NEW can be divided even further into ANCHORED and UNANCHORED. The denotation BRAND-NEW refers to the instantiation of elements for the first time in the discourse. Conversely, UNUSED describes sentence constituents that are readily accessible to the reader, such as *the sun, nature* or proper nouns. The information status ANCHORED is accomplished through a previously established GIVEN entity from within the discourse. A link to previous GIVEN entities is readily achieved through reduced relative clauses and anaphora, as in *the man I know* or *reasons for this*. Finally, UNANCHORED describes (BRAND-)NEW entities that are realized without any link to previous discourse entities, lexicogrammatical, coreferential or otherwise. This is most easily accomplished through indefinite nominal phrases or, in the case of (1b), a null article with a plural nominal phrase as in *laws*. In doing so, a completely NEW entity is realized as a marked NEW theme and contextualized through its instantiation with the GIVEN rheme. However, this is only one piece of the puzzle.

The second factor advanced to account for rhematic regression is found in so-called unmotivated themes, as outlined by Herriman (2011). Such themes can be categorized as NEW or CONTEXTUAL themes. In the case of (1) above, the theme *laws* lacks a referent in the previous clause. In the complete example cited in Shi (2013), there is, in fact, no coreferential element to *laws* in any of the preceding clauses. This forces the theme to be introduced as NEW for coherence to be established between sentences. Cohesion is then established rhematically through the coreference between *the beaver* and *they* in the two clauses. Otherwise, through the collocational relationship between *in danger* and *protect* (and secondarily, *laws*), the thematic *laws* in (1b) qualifies as a CONTEXTUAL theme as well. Regardless of whether *laws* is a CONTEXTUAL or NEW theme, either reinforces the NEW (information) status of the theme in the clause and thereby the claims made above as outlined by Prince (1981) and Kopple (1991).

A final way to account for rhematic regression can be found in the distinction van Dijk (1977) makes between sentence and discourse topics. In (1) and specifically the original text in its entirety, *beaver* is realized continuously as the theme in previous clauses and becomes GIVEN. This promotes the term and concept of *beaver* to a discourse topic within the text. The theme *beaver* thereby informs the informational and discursive global structure of the text. Even when instantiated rhematically in (1b), *beaver* retains its GIVEN status on account of its discourse topic status. Conversely, *laws* represents the sentence topic, which van Dijk describes as a semi-topic, "[…] where the topic not simply entailed by previous discourse still conveys NEW information" (1977: 59). The NEW theme *laws* may have been inferred or entailed within the frame *danger* introduced in the previous clause, van Dijk argues further, which would again merit this deviation in thematic, but not rhematic, progression.

Such inversion of information status via rhematic regression and semi-topics can mark a contrastive shift in previously established discourse topics. In (1b), the use of *but* explicitly introduces a cohesive contrast to what was stated previously in (1a). Together with inverted information status, the author places particular emphasis on *laws* within the GIVEN context of *beavers*. This is similar to how elements can be fronted in written and spoken speech to increase their saliency, e.g., *The queen, I don't know*. As the cohesive tie *but* and the entailment between *in danger* and *laws […] to protect* are present, they further aid in the progression across both sentences. Had these been absent, then a break in the rhetoric would have emerged and thereby a break in the logical development of the discourse message.

Inspired by rhematic regression, Hawes & Thomas (1997a) and Dou & Zhou (2018) conceptualized a further classification of how rhematic elements can be developed as

subsequent rhemes. Known as **rheme reiteration** or **constant rheme**, a rhematic element is repeated across two concomitant clauses, as illustrated in the following example from Dou and Zhao (2018: 64):

(2a) Jim **likes playing basketball** (R_1).
(2b) And Tom also **likes playing basketball** ($R_1 \rightarrow R_2$).

In this instance, an exact lexical repetition of the rheme *likes playing basketball* occurs in both clauses. Repetition of the identical lexical item in concomitant rhemes is not a requirement, however; lexically derived forms of the rheme, such as *sport* as a superordinate term for *basketball* in (2), are also appropriate. Importantly, the element expected to change in informational content, the rheme, remains constant whereas the theme changes. As Hawes argues, keeping the rheme static fulfills emphatic rhetoric in writing (2010b: 49): Structural parallelism through lexical repetition, akin to Hawes' constant type progression, makes rhematic elements GIVEN; developing the foundational theme with a new discourse topic affords it NEW and thereby marked status. Hence, the afforded markedness allows the theme to stand out rhetorically against the repeated rheme and simplifies both text structure and content.

The order of the sentence constituents in (2b) could be switched to instantiate *basketball* as the theme through simple linear progression. Doing so would potentially necessitate the passive formulation of (2b), which would result in marked word order at the expense of the clauses becoming cumbersome. Further, the emphatic rhetoric expressed through constant rheme progression would disappear.

A point of critique against this structure is through Hawes' own constant type progression. In (2), the Themes *Jim* and *Tom* are both proper nouns, which reflect parallelism by means of word class. Therefore, in instances where constant type progression can be proven, it may assume priority. Where thematic parallelism is absent but rhematic reiteration evident, then the latter comes to define the progression in the sentence pair. Ultimately, the benefit to rheme reiteration was already alluded to through the parallel structure of both clauses: it affords stronger cohesion and comprehensibility through lexical repetition, and achieves greater coherence through the discourse topics in the reiterated rhemes.

While the aforementioned explanations do indeed account for rhematic progression, it is critical to note that this constitutes an exceptional progression pattern. After all, it is the rheme, not the theme, that forms the destination of the progression. This stands in stark contrast to all other forms of thematic progression, whereby, as the name suggest, the theme is the destination for reinstantiated thematic or rhematic elements. It therefore goes against the originally postulated theory that thematic progression establishes the structural framework within which a text is developed. For this reason, most researchers forego rhematic progression and opt for a thematic break instead.

In fact, assuming rhematic progression in place of a thematic break could undermine the author's intention of causing a break in the rhetoric at a given point. If the author deliberately broke thematic progression to emphasize or contrast propositional content, attempting to account for discourse development through rhematic regression could alter how the text is received. This, in turn, could result in an incorrect association between the salience of discourse topics and their discursive treatment at that point in the text. It is for this reason that rhematic progression patterns remain controversial. That being said, where plausibly accounted for, they can highlight structural and logical development where the original thematic progression

patterns would fail. Their inclusion as a potential progression pattern therefore has the potential to provide more fine-grained insight into how the text unfolds structurally and deliberately.

3.2 Method of Development and Text Structure

In the previous sections, thematic and rhematic progression patterns were shown to indicate the diverse ways of developing the foundation (theme) or core (rheme) of a discourse message. The collective interplay between these two constituents manifests through a text's unfolding and reveals deliberate scaffolding – deliberate thematic selection – behind the text. As Scarpa states, "[…] the patterning of thematic selection in a text is the thematic progression of that text, i.e., the ways the theme and rheme in a clause connect to those of surrounding clauses as the text unfolds" (2020: 33). Thus, tracing how themes and rhemes are realized at the various structural levels of a text by means of thematic progression can provide insight into how a text is structured on the whole.

Clauses form the microstructure of a text and are the building blocks for stringing together themes and rhemes through thematic progression. The collective clustering of clauses then forms the text's larger macrostructure, the paragraph, whose discourse topic or hypertheme informs its themes and rhemes (Daneš 1976: 109). The conjunction of paragraphs ultimately constitutes the entirety of the text's macrostructure, again motivated by the individual hyperthemes. However, while hyperthemes emerge at the paragraph level, so-called macrothemes underlie the text on the whole, as originally postulated by Martin (1992). These function as the overarching, key discourse topics of a text and inform the resulting hyperthemes.

The structural hierarchy of thematic instantiation from the overall text to the clause is thus macrothemes > hyperthemes > themes (Martin 1992: 443). It is this hierarchical division amongst macrotheme, hypertheme and theme that then embodies the breakdown of discourse topics from general to specific, from abstract to concrete. In other words, the discourse topics at superordinate levels inform the selection of discourse topics at the subordinate levels. What ensues is a parallel system of discourse development on the one hand and structural development on the other. These developments, in turn, inform a text's overall thematic progression.

At its core, thematic progression indicates how the language user consciously selects discourse topics and develops them either as the foundation or the core of a discourse message. Tracing these thematic choices reveals what Fries calls the method of development (1995: 323), whose theoretical framework was also present in Matthiessen (1995) and Lemke (1983, 1994). In his definition, Fries states that "the experiential content of Themes correlates with what is perceived to be the method of development of a text or text segment" (1995: 325). Since topical themes form the foundation of a discourse message, identifying how they are developed – either through the same foundational themes or NEW rhemes – illustrates a text's method of development. Briefly, it shows "the way in which a text develops its ideas" (Fries 1995: 323). Should marked themes occur together with topical ones, then an overt shift in development may be at hand. In other words, the use of textual, circumstantial or interpersonal themes can signal both contextualization of the discourse in the resulting sentence and potential shifts in structural, logical or rhetorical development.

To illustrate a text's method of development more concretely, a problem-solution text has been provided and diagrammed below in terms of its thematic progression. The text was taken from an online writing resource tool that explains and exemplifies the purpose of problem-solution texts (Morton 2011). Themes are in **bold**, pertinent rhematic elements in *italics* and marked

themes underlined. Only sentence constituents that contribute to thematic progression across sentences have been marked for clarity's sake. Note that marked themes are part of the same thematic constituent structure and are considered marked due to their realization before the topical theme.

1. (a) **It seems like there** (T_1) has been a surge in *teen pregnancies* (R_1) these days.
   (b) **Teen pregnancies** (R_1→ T_2) make it very *difficult for young mothers to pursue their dreams and meet the demands* (R_2) of an infant.
   (c) **Fortunately, most teen pregnancies** (T_2→ T_3) can be easily prevented by using *birth control* (R_3);
   (d) **however, even birth control** (R_3→ T_4) is *not 100% effective* (R_4).
   (e) **The most effective way to prevent teen pregnancies** (R_4→ T_5) is *abstinence* (R_5), which is 100% effective.

a) T_1: Dummy-*it* and existential	R_1: Teen pregnancies
b) R_1 → T_2: Teen pregnancies	R_2: Consequences of teen pregnancies
c) T_2 → T_3: Teen pregnancies	R_3: Pregnancy prevention: birth control
d) R_3 → T_4: Birth control	R_4: (Lacking) Birth control effectiveness
e) R_4 → T_5: Most effective way	R_5: Abstinence

The first sentence introduces the problem of teen pregnancies in the rheme through the thematic dummy-*it* and existential *there*. Both structures emphasize the NEW information status of the problem but are not a requirement for the first sentence of a text. The author could have presented the problem as the foundation of the message in the theme since all elements in the first sentence of a text are always NEW. In sentence (1b), the previously rhematic *teen pregnancies* is then instantiated as the theme through simple linear progression (NEW to GIVEN progression) to reinforce the problems that accompany teen pregnancies. In (1c), however, a shift in the rhetoric occurs: the author moves from the problem of teen pregnancies to a possible solution, birth control. This shift in method of development is initiated through the marked modal theme *fortunately* together with constant progression with the thematic *teen pregnancies*. The NEW rheme *birth control* from (1c) becomes the foundation of the message in the theme of (1d) through simple linear progression again to develop this discourse topic, albeit with another shift in rhetoric. This shift is marked through the textual theme *however* and serves to further accentuate the transition to the solution portion of the text. It also serves as a means of texture between sentences (1c) and (1d). The text concludes by instantiating the rhematic *not 100% effective* in the theme of (1e) through lexical repetition. The hypertheme *teen pregnancy* realized together with the thematized *effective* highlights the return to the overarching discourse topic of teen pregnancies as a problem with its proposed solution, abstinence.[13]

The takeaway from this example, as originally touched upon in Fries (1995: 336), is that thematic selection, i.e., the deliberate choice of certain themes, mirrors the presentation and flow of information required to fulfill a text's discourse goal and rhetorical organization. As shown through the modal and textual themes *fortunately* and *however*, marked themes can often indicate shifts in the method of development as progression to a new rhetorical section of the

[13] It should be noted that multiple progression patterns can be present across sentences. For example, while simple linear progression is evident between (1d) and (1e), this sentence cluster also includes gapped continuous progression through the thematic repetition of *teen pregnancies*. As this is the primary discourse topic of the text, some may argue that gapped continuous should be given priority at this point over the rhematic discourse topic *effective(ness)* in (1d). Both patterns successfully contributed to the maintenance of topic coherence and cohesion across sentence boundaries, however. The importance of thematic patterns in text development is less so their type (e.g., simple linear progression) and more so their effect on the structuring of discourse.

text. As Downing & Locke found, “dependent clauses in initial position provide meaningful frames within which the rest of the sentence develops” (2006: 236). This extends to sentence-initial adverbials and phrases as well, which are dependent in nature (Biber et al. 2007: 772).

The greater conclusion to be drawn here takes the harmony between thematic progression and method of development a step further: Method of development is made manifest through thematic selection, which, in turn, can be traced through thematic progression. Whereas the former provides insight into the macrostructure of a text, the latter illuminates the microstructure. Method of development, thematic selection and thematic progression are thus wholly intertwined, which explains why failure to effectively uphold one often leads to failure of the other. A lack of cohesion across sentences on the grounds of absent coreferential elements, either thematically or rhematically, may lead to the absence of overall coherence at the micro-level. Instantiating thematic elements that have not been derived from previously established discourse topics (themes, rhemes, hyperthemes or macrothemes) can equally cause breaks in coherence and discourse. Ensuring that themes can be traced back throughout a text’s method of development is therefore pivotal in preventing structural or logical breaks in the discourse.

That being said, deliberate breaks in the discourse can also be part of a text’s method of development, particularly for rhetorical purposes. In the example above, marked themes were employed to indicate a shift in the rhetoric; breaks, too, can fulfill this function. The most overt kind of thematic break would be transitioning from one chapter in a book to a new one. While discourse themes may carry over, this is by no means a requirement. Within self-contained texts such as paragraphs, thematic breaks can be used for emphasis or to intentionally give the reader a pause. As Hawes demonstrated:

> [B]reaks lend themselves readily to exploitation in evaluative rhetoric and present a particularly interesting category. They are often used to change the direction of the discourse and typically occur at the beginning or end of a rhetorical segment, breaking the flow of thematic progression and thereby revealing the seams of a writer’s ideological message. (Hawes 2010a: 4)

Even Daneš mentioned in his original explication of thematic progression that clauses may intentionally have an “omitted link” or “thematic jump” (1976: 121), which he later substantiated by stating that “[…] there cannot be right and wrong places to break but only more appropriate or less so, depending on the rhetorical needs of the writer” (1995: 30). For that reason, while conscious and conscientious selection of particular themes motivates thematic progression patterns for rhetorical and discursive purposes, intentionally breaking thematic patterns to indicate discourse and rhetorical boundaries can be an equally effective tool.

To conclude this section and tie in the major notes from Chapter 3.1 on thematic progression, the theme, whether marked or unmarked, is the crux of a communicative message at both the micro- and macro-scale. Its discursive function as the foundation of the message suggests its indispensability in terms of how a text unfolds. While the rhematic discourse goal brings the communicative message further, previously established themes form the cornerstone of newly introduced discourse topics.

The degree to which a message is successfully developed can thereby be traced back to how effectively themes are employed throughout the discourse. Where thematic scaffolding of a message is lacking, absent or insufficiently developed, communication runs the risk of breaking down due to gaps in logic, cohesion or information. Through the interplay between theme and rheme as a manifestation of a text’s thematic progression, the underlying discursive functions

and method of development come to light. The marriage of thematic progression, discursive function and method of development thus facilitates the discourse message and potential conventions characteristic of texts and communication.

3.3 Thematic Progression, Genre and Text Type

How a text is realized lexicogrammatically and cohesively is underpinned by the continuous selection of themes as the foundation of communication development. Ultimately, this forms the overall thematic progression structures throughout a text, revealing its micro- and macrostructures, rhetorical functions and discourse goals. These three aspects are intertwined and, critically, mutually realizable, as highlighted in the following quote by Bakhtin (2010):

> All the diverse areas of human activity involve the use of language. Quite understandably, the nature of forms of this use are just as diverse as are the areas of human activity […] Language is realised in the form of individual concrete utterances (oral and written) by participants in the various areas of human activity. The utterances reflect the specific conditions and goals of each such area not only through their (thematic) content and linguistic style, that is the selection of the lexical, phraseological, and grammatical resources of the language, but above all through their compositional structure. All three of these aspects – thematic content, style, and compositional structure – are inseparably linked to the *whole* of the utterance and are equally determined by the specific nature of the particular sphere of communication. Each separate utterance is individual, of course, but each sphere in which language is used develops its own *relatively stable types* of these utterances. These we may call *speech genres*. (Bakhtin 2010: 60; italics from original)

Yet, this tripartite system is one that in actuality is more motivated than what a text's mere surface structure may allude to. What drives this motivation is the concept of **genre** and the closely related **text type**. The present section aims to identify the interdependent (realizational) roles that genre, text type, context, discursive purpose, thematic selection and thematic progression have in a text. This serves as a first step towards illuminating whether thematic progression is emblematic of text type or genre.

Finding a uniform definition of genre, from a linguistic perspective, has historically proven quite difficult, not least because of the myriad concepts closely related to this term. Most research, while theoretically demonstrative, has led to a marred understanding of what genre specifically is and how it differs from related terms, such as text type and register. The first task in attempting to elucidate a possible connection between genre and thematic progression is therefore establishing a clear definition of genre (for an in-depth and substantiated discussion of dissecting genre, register, and text type, see Lee 2001).

Beginning with researchers who contributed significantly to a modern understanding of genre, Martin defines this as "a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture" (1984: 25). He later expounds on this definition by describing it as "how things get done, when language is used to accomplish them [language activities]" (Martin 1985: 248).

According to Kress, genre revolves around

> […] the structural features of the specific social occasion in which the text has been produced, [seeing] these as giving rise to particular configurations of linguistics factors in the text which are realizations of, or reflect, these social relations and structures. (Kress 2014: 33)

Halliday & Hasan (1976) note that considering the analytical factors of context and language use under the umbrella of genre is "[…] not an interpretation of what text means, it is an explanation of why and how a text means what it does" (1976: 378). Biber states that genre can be identified through traditional criteria that stem from text-external, non-linguistic considerations (1995: 70). These "traditional" criteria are those such as "intended audience, purpose, and activity type" (Lee 2001: 38). They aid in forming a category for the given text which has been shaped by the culture and use (Biber 1995: 170). Swales claims that "genres are 'owned' (and, to varying extents, policed) by particular discourse communities" (1990: 24–27).

In more recent interpretations and delineations of genre, Eggins argues that text analysis with respect to genre must consider schematic structure, register configuration and individual realizational patterns employed within the discourse (2004: 56–57). Burlaga bases a renewed understanding of genre on a social constructionist approach, stating that "[genre] is no longer viewed as fixed and immutable but as purposeful and constructed in response to social contexts (2004: 41). Further, Badger defines genre as "a purposive language event" as a result of harmonizing a text's "purpose, function, [and] goal" (2003: 257). Vian Jr. & Lima-Lòpes offer a more robust definition as derived from the context of culture:

> When analysing the context of culture, we should try to describe how the interaction's general purpose leads us to organize a text in stages. A genre, thus, is structured in stages […] and consists of a social process oriented towards a goal – teleologically oriented, therefore – organized and realized by the register. (Vian Jr. & Lima-Lòpes 2005: 31–32)

Finally, Lee identifies genre as "used when we view the text as a member of a category: a culturally recognized artifact, a grouping of texts according to some conventionally recognized criteria, a grouping according to purposive goals, culturally defined, [and] incorporates a critical linguistic (ideological) perspective" (2001: 46).

Glossing over these definitions, the terms used repeatedly to describe genre are first and foremost social and cultural context, discourse purpose and members, and structure. These aspects particularly suit a systemic and functional understanding of language but are not limited to these ideological approaches. Genre can thus be seen as both an abstract and concrete system: abstract, in the sense that it functions as a reflection of the collective ideals, expectations and standards of a social context; concrete, in that a genre expresses actual realizations of these functions as defined by the social context. The variation and variety in language use depend on a social context's discourse purpose and generally accepted criteria from discourse members belonging to a particular context of culture. As these language instantiations begin to reflect systematic similarities in their structure, discourse goals, and group membership, they become conventionally grouped together under a genre.

Culture – including its members, conventions and social practices – is thus a greater foundational construct which both molds and reflects the characterization of interaction among its respective members. It represents the highest level of abstraction in this model, from which

genre stems and which genre embodies. With genre, it is noteworthy that it has less to do with the actual linguistic, i.e., lexicogrammatical, patterns and much more with the extra- or non-linguistic features related to the genre. Discursive context, discourse members, (textual) membership, and discourse structure and goal all lie at the center of genre's definition. That does not mean, however, that specific language patterns are not characteristic of a particular genre. Simply uttering *Once upon a time* would evoke for many its association with the narrative genre, as would the concluding phrase *And they lived happily ever after*. Through conventionalization, these phrases have come to be part and parcel of a specific narrative genre and rhetorically serve to begin and conclude the telling of a fairytale story. Such language patterns are not part of the definition of genre but are instead covered by the system of register and text type, as discussed below. By mapping rhetorical or discursive functions, e.g., how to conclude a fairytale, onto the phrase through conventionalized use over time, a genre and thereby genre-specific criteria emerge.

These functions are further reflected in the schematic structuring of a text, in other words, how utterances are organized according to the stage in the structural development of a text. Tracing the development of rhetorical functions through their linguistic realizations thus falls under text type and register analysis, to which thematic progression belongs. After all, thematic progression itself functions as a tool for dissecting a text's rhetorical and logical structure through thematic selection. Before considering the viability of this statement, however, it is necessary to outline what text type is, how it differs from genre and what its importance in text analysis is.

Aumüller (2014) provides a succinct yet meaningful explanation of the distinction between genre and text type:

> While genres single out entire texts according to heterogeneous features (e.g., formal features in the case of sonnets and paratextual information in the case of homepages), text types try to capture semantic relations between textual surface structures (of items of *discours*) and content structures (of items of *histoire*). (Aumüller 2014: 857)

Simply, text types are a narrower instantiation of genre, whereby certain linguistic features, semantic or textual, are further delineated for a specific discourse purpose in a specific social context. For instance, the narrative genre not only entails novels but also fan fiction, short stories and comics. These text types, in turn, can be refined even further depending on the representative and characteristically recurring linguistic features. These linguistic features, extending to the extra-linguistic, are the focal point of investigation when establishing potentially conventionalized or characteristic patterns in a particular text: What cohesive ties are most commonly used in the text? Is the rhetorical quality of the text more informational, argumentative, presentational or interpersonal? How is the text visually, semantically and textually structured? Are the same patterns and features found intertextually, i.e., in other texts? Uncovering overt and underlying semantic, discourse and textual patterns allows determining the pervasiveness of certain norms within and across text types. In turn, text type categorization facilitates attributing more or less uniform text types to specific genres. Ultimately, however, this begs the question: to what end?

Characterizing a text in terms of its texture features' frequencies alone may have its scholastic merits but otherwise bears little fruit. For example, knowing that rhematic progression is a comparatively rare form of structuring a scientific text may, at first glance, appear to be of little importance. However, as preliminary findings from the present research revealed, this progression patterns introduce researchers as NEW in the theme and their research as a GIVEN

rheme, which runs contrary to the conventional theme-GIVEN, rheme-NEW paradigm (for instance, *Howes (2020) showed that texture patterning is irrelevant under certain conditions*). Hence, a contrary structuring in informational status emphasizes elevates the researchers as both the grammatical subject and as a NEW discourse topic; their research then forms the GIVEN discourse topic, which is further expounded upon in the accompanying rheme. This patterning demonstrates a micro-shift at that stage of the text's development and becomes a characteristic of scientific writing through recurring use (Hawes 2015: 98). By mapping texture features to rhetorical functions in a text type, the effect that texture features have as a discourse tool comes to light.

Employing rhetorical and linguistic features emblematic of a text type readily allows readers to identify a text as such and thereby establishes a degree of expectation vis-à-vis the communicative purpose of the text (Figueiredo 2010: 130). This then facilitates a better understanding of the structural makeup and discourse goal of a text as a guide for the reader. Remember that genres and text types are a reflection of culturally accepted conventions, linguistic or otherwise, belonging to that category of text. Their use thus provides insights into the social and discourse members of the text type. Examples of this are using group-specific lexis (terminology originally stemming from and employed by minority groups), employing formatting patterns (the use or omission of punctuation in text messages depending on generation), and lexical density. As shown in Chapter 3.2 on method of development, the texture characteristics of a text type are an amalgamation of structure, rhetorical function as well as social and discourse context. However, researchers have yet to come to a clear consensus as to whether genres or text types reflect unique structural patterning via thematic progression. In other words, while some argue for a correlation between text type and thematic progression patterns, others are less convinced of this claim. Research highlighting this potential relation will be presented in the following in order to shed light on both sides of the argument more closely.

Table 3-1 in the following summarizes the relevant contemporary work undertaken on thematic progression and text type, and provides a strong foundation for those interested in delving into the topic. Researchers who purport a positive correlation between thematic progression and text type are Halliday & Hasan (1976), Martin (1992, 1993, 1995), Fries (1981, 1992, 1995), Francis (1989), Swales (1990), Berry (1995), Matthiessen (1995) and Hawes & Thomas (1997b).

It should be noted that the these authors do not attribute thematic progression patterns alone to the identification of a text type or genre. Rather, it is one of the numerous factors that aid in qualifying a text as belonging to a certain type or genre. The rhetorical functions and associated discourse goals of a text are equally important to the overall textual and communicative structure. Again, the harmony between these elements and their reciprocating realizational factors are pivotal in their reflection of text type characteristics and membership.

Thematic progression patterns as an important albeit functionally inferior factor in determining text type membership is a position held by researchers such as Loftipour-Saedi & Rezai-Tajani (1996), Mauranen (1993), Sinclair & Carter (2014) and Winter (1982). The arguments put forward by researchers in this camp recognize the function of a text's thematic selection; however, they argue that thematic progression patterns are secondary to the rhetorical functions and communicative goals of a text. Furthermore, these researchers claim to have found little evidence to support the theory that thematic progression patterns themselves are unique to specific text types or genres. According to these researchers, the same thematic progression patterns can be found across disparate text types and genres and at relatively equal frequencies. In other words, similar frequencies of thematic progression patterns and therefore thematic

structure regardless of text type are incapable of representing a text type. Instead, rhetorical functions and other texture factors, such as lexical density and cohesive devices, contribute more to text type membership. Hence, reducing a text to its thematic structure alone would prove insufficient in allocating a text to a specific text type.

Text Type	Researchers
News articles	Burlaga (2004), Gómez (1994), Hoey (2005), Downing (2001)
Scientific research articles and abstracts	Martínez (2003), Lorés (2004), Hasselgård (2020), Leong (2015, 2019), Leong et al. (2018), Ebrahimi (2016)
Science essays by undergraduate students	North (2005)
Promotional texts	Ho (2017)
News reports on same issue but from different regions	Lu (2002)
Academic writing by non-native English speakers	Jalilifar (2010), Hawes & Thomas (2012), Herriman (2011)
Introductory sections of academic textbooks	Jalilifar & Montazeri (2017)
Biomedical articles	Kanoksilapatham (2007)
Conference papers	Naderi & Koohestanian (2014)
Editorial articles	Hawes (2010b)
Magazine editorials	Figueiredo (2010)
Tabloids	Hawes (2010b)
Children's storybooks	Guijarro & Zamorano (2009)
Expository and narrative genre	Shi (2013)
Undergraduate theses	Gunawan & Aziza (2017)

Table 3-1: Contemporary research conducted on the correlation between thematic progression patterns and text type from the last 20 years.

Despite these claims, a greater wealth of research, as summarized in Table 3-1, tends to indicate that thematic progression should not be treated secondarily. With due consideration of the thematic progression patterns behind the text, the communicative goals as realized through rhetorical functions in their staging are made more prevalent and purposeful. A text's discourse development can thus be brought to the forefront through the analysis of its thematic selection and progression. Taken together with other texture characteristics, thematic patterns help to shape the final form a text may have as a singular example of its greater text type and genre.

3.4 Computational Approaches to Thematic Theory

All research on thematic progression presented thus far shares a common methodological approach in that analyses were performed manually, i.e., by hand. This approach was the only option before the advent of natural language processing (NLP) and corpus-based methodologies. However, even nowadays, only limited research has been dedicated to an automated, computational analysis of thematic progression. As current trends in linguistic research are experiencing a greater use of natural language processing, the present section will outline the relevant research conducted on the automation of thematic progression analysis to date.

Over the previous two decades, there has been a considerable increase in both automated and computational analysis of language on account of the expansion of computational capabilities (Khurana et al. 2022). The strength behind such an approach is the ability to analyze vast

datasets that a manual analysis categorically cannot achieve. Most research has revolved around automated text segmentation or automated topic identification for summary purposes (Popping 2000; Leopold & Kindermann 2002, Hotho et al. 2003, Wiedemann 2016, Singh et al. 2010, van Atteveldt et al. 2021). When exploring the scope of research conducted on automated thematic analyses, it at first appears as if this is a richly researched field of study. However, upon closer examination, 'thematic' in these studies is used synonymously with the term 'topic' or 'gist' and not in the sense of thematic theory (see, for example, Scharkow 2013, Lancia 2012, Boyatzis 2010).

That being said, studies exploring the computer-aided automation of theme identification have begun to increase, such as the work done by Lavid (2000), Moens (2007), Hajičová & Mírovský (2018), Leong (2019), and Xi et al. (2020). In fact, research has shown that thematic progression as a text analysis parameter can improve algorithmic efficiency in machine translation, information extraction, information retrieval, and text classification (Steinberger & Bennett 1994, Kappagoda 2009). Three particular groups of researchers that furthered the thematic paradigm from a computational linguistics perspective are Schwarz et al. (2008), Park & Lu (2015) and Domínguez et al. (2020). Since the present study took inspiration from their work, the following will elucidate in more detail the methodological background and findings from their research efforts.

Schwarz et al. (2008) developed a rule-based system for automatic theme identification, whereby a text was initially parsed with the Stanford PCFG parser. Their tool then produced output in terms of the text's parts of speech, phrase structure and dependency structure. The model followed a Hallidayan approach by identifying themes based on 85 sample sentences (Halliday & Matthiessen 2004: 65–81). From these, 14 rules for theme identification were ultimately derived. The resulting parser from Schwarz et al. was a tree-rule parser that analyzed a text by means of these rules to output a parse read from a parse file. The actual automation was achieved through a Java application and requisite XML file, which was the collection of rules to be applied to the parse file.

Their algorithm was tested on two sets of data, a smaller sample of four texts with 48 sentences total and a larger corpus consisting of 209 academic abstracts amounting to 700 sentences total. While the accuracy of the smaller sample texts was lower (83.33% of sentences were classified with a precision of 59.74%), the results from the larger corpus achieved an accuracy rate of 81.74% precision upon classifying 89.28% of sentences (Schwarz et al. 2008: 22). The precision of the tree-rule parser was largely affected by novel theme types. Since a number of sentences included multiple themes, the parser was unable to account for these correctly. This was ultimately to the detriment of the smaller sample and its precision rate. The higher accuracy for the large corpus indicates the lower frequency of multiple themes in the sampled sentences. The research by Schwarz et al. represents the initial foray into the automated constituent analysis of a text's themes. While thematic progression was not considered, the research laid the foundation for subsequent research in the field.

The second research group, Park & Lu (2015), built on the rule-based system by Schwarz et al. (2008). Their work employed machine learning to develop the software Theme Analyzer for the automatic identification of the theme in a t-unit. This basic analytical unit included the matrix clause plus any hypotactic and/or subordinate clause(s). The researchers followed a systemic functional grammar approach and, in contrast to Schwarz et al. (2008), readily accounted for multiple themes within a single sentence. In addition to requisite topical themes, optional textual and interpersonal themes were therefore embedded in the parsing functionality of this model. The first topical element of a sentence was delineated as the theme; everything

afterwards was then marked as the rheme. In addition to automatically identifying the theme, Park & Lu's application also determined the theme boundary, the theme's syntactic nodes, function (topical, textual or interpersonal), markedness, role (subject, complement or adjunct) and finally the sentence mood type.

Altogether, Theme Analyzer used 264 theme-tagged model sentences to test 250 expert sentences from the *Wallstreet Journal* section of the Penn Treebank and 250 student-level sentences from the British Academic Written English Corpus. Their program achieved an accuracy of 93.0% in correctly identifying the theme(s) in a t-unit. Additionally, Park & Lu found that expert writers more frequently employ a wider range of syntactic nodes and marked themes, which suggests "a wider repertoire of thematic choices" than student writers (2015: 97).

The final piece of research that not only automated theme identification but also thematic progression is that of Domínguez et al. (2020) and their software ThemePro. Their work employed the Spacy parser, which is a Python-based Application Programming Interface (API) that uses machine learning and trained language models for natural language processing applications. The parser was used specifically to extract universal dependencies from the text input, which were then converted to the CoNLL format for subsequent linguistic analyses. This format allowed the creation of tree structures on the basis of the textual input, which was traversed together with pattern-based rules to determine the requisite themes. For thematic progression, the so-called cosine similarity, a numerical representation of the semantic similarity between thematic and rhematic elements, determined the thematic progression patterns present across sentences. Once analyzed, the text's thematic progression was visualized using a web interface and a node network for each sentence. This visualization was provided as a means to indicate the development of thematic elements throughout the text. Other aspects visualized in the output were co-reference chains and syntactic trees.

The corpus used to evaluate their tool was comprised of three stories tallying 1312 words in total. Overall, ThemePro achieved an average classification accuracy of 60%, which was an increase of over 10% compared to that of Bohnet et al. (2013), whose work served as their comparative baseline. Visualization further enhanced the comprehensibility of the results and how thematic elements contributed to the informational development of the discourse. Access to the tool via a web interface partially allowed the use of the tool without a specific operating system. As such, a wider user base was achieved through the tool's public access.

These three pieces of research highlight the cumulative development that the automation of thematic analysis has undergone over the past 20 years. Despite the ever-increasing parsing accuracy that the respective tools achieved, a number of deficiencies were present in their methodologies. Firstly, each research was limited by the scope of the corpora that formed the testing basis of the respective models. This limited the statistical representativity of their results and applicability to text types outside of those used for testing purposes. Further, the parsers were only able to analyze one text at a time, which made intertextual analyses cumbersome, if not impossible. Output from each text would need to be compared to each other manually, which thereby limits the automated analysis the tools afford. Thirdly, annotated corpora were not made available, which prevents the reproducibility of the researcher's results. Currently, there are no publicly available corpora that contain annotated text with respect to themes and thematic progression. Had these been made available, then they could have been used as a foundation for expansionary work on the automation of thematic progression analysis. Finally, while the visualization in the work from Domínguez et al. (2020) provides users with greater insight into the syntactic and informational structure of a text, the output for thematic progression specifically remains nebulous. All themes and rhemes from the text are merely

denoted as a node in a network with their semantic similarity values as connections between the nodes. The result is a collection of the repeated terms *theme* and *rheme* alone without actual reference to the original text. In other words, which theme refers to which part of which sentence in the text is not provided. Therefore, visualization for thematic progression specifically simply became an abstraction of the text in terms of semantic similarity values alone.

Both the contributions these three pieces of work provided and their accompanying shortcomings inspired the present research on an automated approach to thematic progression. Specifically making the visualization of results more comprehensible and accessible to users regardless of background formed a foundational pillar of the software developed for this work. Expanding the analytical functionality of the tool to any number of texts and text types further defined the progressive development from previous research. As elaborated in Chapter 4, similarities between previous and the present work will become apparent in its dependency-based approach to thematic parsing with the application programming interface Spacy and the programming language Python.

3.5 Summary of Thematic Progression as a Tool for Text Structure and Method of Development

The research and accompanying theories presented here in Chapter 3 show that thematic selection and thematic progression are defined by tracing the development of a communicative message through a text's themes and rhemes. As originally put forward by Daneš and further refined in contemporary research, thematic progression reveals the conscious structuring of both propositional and informational content in discourse. In time, contemporary additions to these patterns were posited to make up for deficiencies identified in Daneš's original four patterns. As such, a wider spread of thematic patterning was able to account for the nearly limitless ways sentences can be constructed to further discourse in written text.

Thematic progression was then explored in its relationship to genre and text type. Just as thematic progression patterns are indicative of the micro- and macrostructure of a text, the frequency of these patterns can shape the texture and rhetorical characteristics of specific genres and text types. While much contemporary research found a direct relationship between thematic progression pattern frequency and text type, many researchers remain unconvinced of thematic progression's characterization of a text type. Regardless, the mapping of thematic progression patterns with a text's rhetorical functions and discourse goals remained a joint consensus amongst researchers.

Finally, the chapter concluded with a discussion on computational methods for automating the identification of thematic constituents and thematic progression in text. The research presented therein leveraged statistical, rule-based and machine learning approaches to automated text analysis. The work built upon the theoretical models that have been developed since the initial conceptualization of thematic theory throughout the 20th century while incorporating natural language processing methodologies that emerged in the 21st century. Despite limitations, the research laid the groundwork for a contemporary treatment of the thematic paradigm by computational means.

Chapter 4 – Methodology

This chapter outlines how text analyses are performed programmatically with the software developed in the present work, Thematizer. The present methodological approach was inspired by both the theoretical models presented in Chapters 2 and 3 as well as the following two core research questions: Firstly, how can the present research enrich a contemporary understanding of the thematic paradigm by building upon and overcoming deficiencies in previous models of thematic theory? Secondly, how can thematic theory be operationalized by computational means in order to make it accessible to writers? The motivation behind these research questions and how they informed the direction and goal of the presentation research thereby form the foundation of each section in this chapter.

Chapters 4.1 and 4.2 revolve around the two core research questions outlined previously. First, the theoretical shortcomings identified in previous research on thematic theory are presented. Deviations from the conventional approach to thematic theory, in particular from the Hallidayan perspective, are outlined with the reasoning for a divergent approach. Here, the present work's treatment of marked and unmarked themes serves to partially answer the first research question.

Chapter 4.2 considers the second research question in terms of how previous automated tools approached the thematic analysis of text via computational means. Deficiencies in previous software are outlined as a steppingstone to how Thematizer was developed to overcome such theoretical, programmatic and functional shortcomings. This discussion then highlights both the impetus behind developing the software and the question of how Thematizer operationalizes thematic theory as a function of accessibility to writers.

Chapters 4.3 and 4.4 define the theoretical framework for theme, rheme and thematic progression that the current work subscribes to. Exact definitions, structural analyses and examples are provided to illustrate the thematic models used for this research. These span a breakdown of marked themes, unmarked themes, rhemes and each constituent's role in thematic progression patterning. The thematic model presented in each section thereby defines the theoretical underpinnings that were translated into computational models for the development of Thematizer.

Chapters 4.5 and 4.6 address the materials and tools used for the development of Thematizer. Chapter 4.5 summarizes the training and validation texts used for training and testing the software. An explanation of the reasoning for the chosen texts and the duration of Thematizer's development stages complete this section's discussion. Then, Chapter 4.6 presents the programming tools, application programming interfaces, libraries and reasoning for their implementation. Here, an explanation of the exact versions of the programming tools in addition to their parsing functionality within Thematizer is provided.

The remaining sections of Chapter 4 detail the functionality and development of Thematizer from a programming perspective and with the help of this work's thematic models. Chapter 4.7 briefly outlines the overall parsing steps that Thematizer progresses through, including text pre-processing. Chapters 4.8 to 4.10 detail the three core parsing tasks that Thematizer performs with each thematic analysis: the identification of theme and rheme spans; marked theme identification and classification; and the identification and classification of thematic progression patterns. Marked theme identification and classification require two parsing steps that employ unique dependency and indexical tests. As such, both are treated individually in Chapters 4.9.1 and 4.9.2. Similarly, as a battery of tests are required for the identification and

classification of thematic progression patterns, each test is treated individually with an explanation of its parsing and classification conditions in Chapter 4.10.1 to Chapter 4.10.5.

Chapter 4.11 presents the web interface developed as the frontend for Thematizer with corresponding screenshots. User input and interaction are outlined for each part of the interface. Analysis output and supplemental explanations as they appear on the interface are also outlined. This section presents Thematizer in its design and operability from a user's perspective alone; no actual results from the tool's analytical output are discussed.

Finally, Chapter 4.12 summarizes the functionality of Thematizer and the underlying methodological facets which were derived from the present work's formalized thematic models and core research questions.

4.1 Deficiencies in Previous Approaches to Thematic Theory

The first research question this work set out to answer was whether any deficiencies in the theoretical understanding of thematic theory existed; where present, the next natural question was how to resolve these deficiencies. In an attempt to answer these questions, a comprehensive review of the theoretical framework for the thematic paradigm was undertaken, as outlined in Chapters 2 and 3. This laid the foundation for a contemporary understanding of thematic analysis while allowing insufficient or deficient aspects thereof to be identified. The first of these concerns itself with the treatment of marked and unmarked themes appearing in the same sentence (compare divergent approaches in Table 4-1).

Sentence 1

Traditional Approach	MARKED THEME					RHEME			
Present Work	MARKED CIRCUMSTANTIAL THEME					UNMARKED THEME		RHEME	
Text	*In*	*children's*	*nursery*	*rhymes*		*this*	*correspondence*	*is*	*intact.*

Sentence 2

Traditional Approach	MARKED THEME 1			MARKED THEME 2		RHEME			
Present Work	MARKED CIRCUMSTANTIAL THEME			MARKED MODAL THEME		UNMARKED THEME		RHEME	
Text	*In*	*adult*	*verse*	*of*	*course*	*it*		*is*	*not.*

Table 4-1: A comparison between the traditional approach to marked theme analysis (Halliday & Matthiessen 2014) and the present work's approach. Whereas Halliday & Matthiessen exclude unmarked themes in the presence of marked themes, the present work accounts for both in thematic analyses. This ensures that previously established discourse topics can fulfill their function of establishing the foundation of a discourse message and can be traced accordingly.

Traditional approaches (cf. Halliday & Matthiessen 2014) only allowed for unmarked themes when a marked theme was not present in the same sentence. If a marked theme began a sentence, then the sentence constituents thereafter were defined as rhematic, as shown in Table 4-1. There, the progression of *this correspondence* in sentence (1) to *it* in sentence (2) occurs in the rheme alone due to the presence of marked themes. Since both constituents are GIVEN discourse topics, however, their information status should be reflected in their thematic status. Instead, their allocation to the rheme implies they are NEW discourse topics. Analyzing the marked theme

alone thereby overlooks the contribution that GIVEN discourse topics after the marked theme provide. This contradicts the definitional function afforded to themes, whereby GIVEN discourse topics form the foundation of a discourse message. Further, it obfuscates re-instantiation of foundational GIVEN discourse topics throughout the text. In other words, tracing which discourse topics are employed as the foundation of the message and which topics develop the discourse (i.e., rhematic elements) becomes blurred.

To circumvent this problem, the present work includes both the marked and the unmarked theme in thematic analyses if both are present in a single sentence. By separating the unmarked theme from the marked theme and rheme, GIVEN discourse topics remain accounted for (cf. the present work's approach outlined in Table 4-1). A text's method of development, i.e., the conscious choice of thematic elements throughout a text, can thereby be more readily identified. Additionally, this reinforces the unmarked theme's discourse function as the foundation of the message together with the marked theme.

Accounting for both marked and unmarked themes within a single sentence can also ensure that the thematic progression across sentences is maintained. Marked themes are typically realized as NEW discourse topics to aid in contextualizing the GIVEN unmarked theme and following rheme. The unmarked theme, placed between a NEW marked theme and a NEW rheme, can thereby serve as the connecting element across sentences. In Table 4-1, constant continuous progression can be ensured through the use of unmarked themes realized with their marked themes. While constant continuous progression would also be evident in the traditional approach due to hyponymy between *children's nursery rhymes* and *adult verse*, this is not always guaranteed. Thus, the present approach to including themes of both markedness types is able to uphold thematic structure in both standard and exceptional cases.

The second shortcoming identified in contemporary theories of thematic structure is the identification and categorization of marked themes. The present work follows a similar understanding to marked theme analysis in that sentence constituents before the grammatical subject are considered marked, regardless of experiential, textual or interpersonal metafunction. In Table 4-1, the sentence-initial prepositional phrases are marked themes since they are both experiential and appear before the grammatical subject. This condition is simplified whereby the grammatical subject of the main clause signifies the border between all marked themes and the rheme, similar to how the transition element acts as the boundary between the theme and rheme in the Prague School approach. The grammatical subject therefore is always the unmarked theme, where present. Using the grammatical subject alone as the border facilitates computational tests and parsing as the subject can simply be extracted via its dependency parse; metafunctional information, such as a constituent being experiential or interpersonal, is not considered and would have required an entirely different computational approach through labeling or semantic testing. Therefore, so long as sentence constituents appear before the grammatical subject of the main clause, they are considered a marked theme.

Further categorization of marked themes then represents an expansion of conventional approaches to thematic analysis that the present work adopts. Instead of denoting marked themes as such alone, these are classified into their functional category of structural, modal, circumstantial, hypotactic and projecting themes. Conventional approaches associated metafunctions with marked themes alone (interpersonal, textual and interpersonal, cf. Halliday & Matthiessen 2014: 107), which served as an inspiration for the aforementioned functional categories of marked themes. Further delineating marked themes into such categories was done to highlight the diversity and frequency of marked theme types in text. Expanding the types of marked themes to include hypotactic themes (e.g., *when, after, because, if*) and projecting

themes (e.g., *it is clear that…*, *what is unanswered is whether…*, *results indicate that…*) offers greater insight into the diverse contextualizing functions marked themes have. What is more, this categorization of marked themes could be used to reveal texture characteristics within and across text types. The prevalence of projecting themes in legal texts, for example, could be a characterizing quality of the text type, which could be captured through a fine-grained analysis of its marked theme usage. Furthermore, a finer distinction between the various marked theme types reflects the wide range of syntactic realization patterns that marked themes instantiate, such as subordinate, adverbial clauses for hypotactic themes and a range of prepositional phrases for circumstantial themes.

Finally, the semantic subclassification of marked themes represents an additional analytical step included in the thematic model employed in this work. Once Thematizer identifies one of the five marked theme types, it then classifies the marked theme into its semantic class, as summarized in Table 4-2.

Marked Theme Type	Possible Semantic Subclasses	Examples
Modal Theme	RESERVATION, INTENSIVE, VALIDATIVE, DEGREE, INTENSIVE, TYPICALITY, EVALUATIVE, EXPRESSIVE, DESIDERATIVE, PRESUMPTIVE	*obviously, unfortunately, classically, in particular, for the most part*
Circumstantial Theme	CAUSAL, MANNER, TEMPORALITY, CONTINGENCY, ANGLE, MATTER, LOCATIVE, ACCOMPANIMENT	*in 2018, as for, without, under, along with*
Structural Theme	EXTENDING ADDITIVE, EXTENDING VARYING, ENHANCING SIMPLE, ENHANCING MANNER, ENHANCING CAUSAL, ENHANCING CONDITIONAL, ELABORATION APPOSITIVE, ELABORATION CLARIFYING	*for one thing, except for, beyond, regardless, next, also, like, similar to, in conclusion*
Hypotactic Theme	CONDITIONAL, TEMPORAL, INFINITIVAL, CONCESSIVE, CAUSAL, MANNER	*when, if, because, after, so long as, whereas*
Projecting Theme	ADJECTIVAL, EXPERIENTIAL, OBJECTIFYING, INTERPERSONAL	*It is clear that..., What the results show is that..., Researchers claim that...*

Table 4-2: The five marked theme types that Thematizer categorizes fronted adjuncts, complements and adverbial phrases into. These are then classified further into the semantic subclass of the identified marked theme based on its use in text.

While the delineation of marked themes' semantic classes has been part of previous manual text linguistics analyses, they have yet to be formally included alongside the automated analysis of marked themes. By expanding the scope of marked themes' analysis to their semantic classes, it is argued that their semantic contribution to the development of the text within and across sentences can be more readily traced. Through specification instead of the catch-all term *marked theme*, users will be able to recognize the marked theme as such, identify the general thematic class it belongs to (e.g., circumstantial or structural), identify typical syntactic realizations of that marked theme class, and finally associate the semanticity that the specific marked theme affords.

In summary, the present work identified two core areas for improvement in the conceptual understanding of thematic theory: capturing marked and unmarked themes realized in the same sentence and further classification of marked themes into their functional and semantic classes. It is posited that the incorporation of these two facets will strengthen Thematizer's output through a multivariate and fine-grained analysis of thematic constituents. In doing so, the development of discourse topics in text can be traced more closely.

4.2 Thematic Structure through a Computational Lens

The second research question to shape this research was how the operationalization of thematic theory can make thematic structure accessible to writers. Here, accessibility is seen as a function of how well Thematizer is able to capture thematic structure through its automated and computational analyses. Applied linguistics research methodologies more frequently leverage computational means to natural language processing nowadays due to increased computational efficiency, access and knowledge (Khurana et al. 2022). As touched upon in Chapter 3.4, research has been directed around translation, text summarization, topic identification and information retrieval. However, only limited work has been done on the automation of text analysis in terms of its thematic structure. Furthermore, the tools that are available nowadays suffer from limited scope in analytical functionality and data visualization. This makes the analytical output less accessible to writers and users, in general, who wish to gain insights into their texts on the basis of thematic theory.

The development of Thematizer as an automated tool for thematic analysis serves to overcome this gap in such a way that it both forwards research on the thematic paradigm and aids users in their understanding of thematic structure in text. The theoretical models that underlie the functional core of Thematizer (cf. Chapters 4.3 and 4.4) represent the linguistic framework for pushing forward contemporary research on the thematic paradigm. The tool Thematizer itself, conversely, represents the interface between the theory and the user's engagement with thematic structure through their texts and the analytical output.

In order to facilitate greater accessibility to thematic structure in writing, a number of key features were identified for inclusion in Thematizer's implementation. These were largely identified through shortcomings from previous approaches to automated thematic analysis, specifically from the tools developed by Schwarz et al. (2008), Park & Lu (2015) and Domínguez et al. (2020).

First of all, previous automated tools limited their thematic analyses to general themes and rhemes alone. Instead of distinguishing between marked and unmarked themes, these were subsumed under the general category of theme.[14] Further, no further classification of marked theme types was provided. This generalization of thematic modeling simplified their analysis but at the expense of fine-grained output. As marked themes can offer additional aid in tracing a text's method of development, particularly with shifts in rhetoric, discourse message contextualization and framing, their absence in the output limited insights into the text's thematic structure.

For that reason, Thematizer was conceptualized from the start to account for both marked and unmarked themes (as initially addressed in Chapter 4.1). Where present, marked themes are categorized into the five marked theme classes outlined in Table 4-2. In so doing, the syntactic and semantic contribution of marked themes could be more readily identified and traced through a text's development. Since marked themes are prognosticated to be a textual characteristic of a given text type, analysis of a text's marked theme frequency could shed light on their contribution to discourse development.

Secondly, previous tools were limited by the number of texts that could be simultaneously analyzed. These only accepted one text at a time, which made intertextual analyses cumbersome.

[14] Domínguez et al. (2020: 1003) only considered paratactic and hypotactic clauses as so-called propositions, which received their own thematic analysis, albeit with the general theme-rheme categorization alone.

Results of individual texts would have to be manually collated after the automated analysis was completed. This not only necessitated additional time in the analysis, it also increased the potential for errors to arise during data collation. Thematizer was therefore equipped with multi-document analysis for the purpose of comparative analyses. Such analyses, particularly within and across text types, are of importance in text linguistics since they can reveal both shared and unique characteristics. All texts that are fed into Thematizer are analyzed simultaneously, and their analytical output is presented both individually and collectively. As such, the user can immediately access the results from individual texts and compare results from multiple documents that were uploaded. Thematizer is also able to analyze any kind of text type and of any length, although greater lengths necessitate greater processing time. Including this functionality was an important developmental aspect of the tool since Thematizer was conceptualized to reach as wide of a text type spectrum as possible. The processability of any text type was meant to further facilitate intertextual analyses for comparative purposes (see Chapter 4.11 for screenshots of sample output of Thematizer's (inter)textual analyses).

Thirdly, data visualization was progressed in the work by Domínguez et al. (2020), whereby thematic analyses were visualized via node networks. Thematic and rhematic constituents were first broken down and connected via nodes at the sentence level; at the text level, thematic progression maps were produced to indicate the thematic interconnectivity of each sentence throughout the text. The output produced was thus an abstraction of a text's thematic structure upon analysis. The drawback to their visualization scheme was the interpretability of the output. Since thematic nodes were systematized within an abstracted network, whose connections were given as semantic similarity values only, the text's overall thematic development became somewhat obtuse. What is more, visualization of thematic progression occurred without its corresponding textual realization and without reference to the actual sentence number. Therefore, it remained unclear which sentence was referenced within the thematic progression network, how it developed thematically from a previous sentence and how it contributed to the text's overall thematic structure.

Comprehensive and comprehensible data visualization formed a key component of Thematizer's output; after all, it is through visualization of the output that the user can understand the underlying analyses and characteristics of their text. Since previous knowledge of the thematic paradigm is not a requirement for Thematizer's use, the output it produces and visualizes should facilitate an understanding of thematic structure on the basis of the user's text. A web interface was therefore developed where the user can upload texts and parse the subsequent results. Instead of producing an abstraction of the thematic structure separate from the user's actual text, the analysis of thematic constituents, marked theme classification, thematic progression and overall frequencies are presented with and within the user's text. Particularly where thematic progression is concerned, the thematic constituents that initiated progression, the means of progression (e.g., lexical repetition or paraphrase) and the progression patterns are embedded in each sentence of the user's text. Explanations of all results with accompanying examples are also provided in the results of the web interface for those interested in a more comprehensive understanding of the output. It is argued that such minutiae in the results and their visualization are critical to making thematic structure both accessible and tangible.

The inability to export the analytical output to the user was identified as a final shortcoming of previous tools. While results were provided in partially visualized form, no option was offered to export them for personal use. For researchers in particular, this presents a limitation to the application of the results since they would have to be manually collated. Considerable time for

manual transcription of the results and the potential for error during transcription further exacerbated this limitation.

To overcome this parsing hurdle, Thematizer was equipped with the ability to save the results in common machine-readable formats: JSON, CSV or Excel. Included in the file are part-of-speech tags, dependency parses, partitioned themes and rhemes, marked theme types and semantic classes, thematic progression patterns, means of progression, progression-instantiating elements and the text's filename. These results form the textual and analytical output of the data that is visualized in the web interface. Being able to export the results gives agency over one's own text analyses as users can then use the analytical output for their own use. Since no publicly available corpora exist that have been annotated specifically for thematic progression, this tool was developed for linguists in particular to facilitate open-access and proprietary thematic progression corpora for their own research. Through greater accessibility to data on thematic structure, this has the potential to expedite, enrich and further research on thematic theory within the linguistic community.

In summary, this section addressed how the operationalization of thematic theory via Thematizer can make thematic structure accessible to writers. Inputs to a potential answer were provided through the initial treatment of Thematizer's core functionality vis-à-vis previous thematic analysis tools. Thematizer's functionalities are thus the present work's solutions to shortcomings of previous computational approaches and to hindrances in making thematic structure accessible to writers. The accuracy with which Thematizer successfully employs thematic theory for automated thematic analysis is the final piece required to conclusively determine Thematizer's degree of successful operationalization. Before this can be answered, however, the materials and methods employed for the development, training and testing of Thematizer will be outlined.

4.3 Definition of Theme & Rheme in the Present Work

Against the backdrop of these research questions, the theoretical framework that this work employs will be outlined in the following in full. As will be explained in the subsequent methodology sections, these theoretical definitions and models served as a bedrock for how Thematizer was programmed. How these theories are put into practice is thus embodied in Thematizer's computational implementation.

The definition and delineation of theme and rheme in this work take inspiration from both the Hallidayan and Prague School approach; however, as already outlined above, deviations from both schools of thought are present. Whereas the Hallidayan approach to thematic theory established the general framework within which the present work rests, the multivariate approach to thematic analysis by the Prague School informed the present treatment of marked themes. Specifically, sentences can be broken down into marked themes, unmarked themes and rhemes, similar to Halliday & Matthiessen (2014). In line with the Prague School, albeit without shared terminology, marked themes are broken down into multiple functional and semantic classes. This is reminiscent of the classification of themes into diatheme, diatheme oriented elements, theme proper and theme proper oriented elements and the semantic function of fronted elements (SETTING and SPECIFICATION, cf. Chapter 2.3). As with both schools of thought, the exception cases of interrogatives, theme-less constructions, existentials, predicated themes, clefts and projecting clauses are also accounted for in the present framework.

Conceptually, the theme is defined as the basis of the communicative message in the text whose propositional content has been previously established within the discourse. Here, 'previously

established' refers to the explicit realization of the discourse topic at a previous point in the text. The theme thereby represents the continuous recapitulation of unfolding discourse topics throughout a text. Conversely, the rheme is defined as the discourse topics that expound on the thematic basis as a means to develop the communicative message further. The rheme aids in achieving the discourse goal by specifying, exemplifying, contextualizing or elucidating the thematic foundation. The discursive function of the rheme is thus to introduce novel discourse topics and information on the basis of the theme.

On account of the context-independent information that the rheme presents, i.e., discourse topics introduced for the first time in text, the rheme most commonly has NEW information status. The theme, then, largely has GIVEN information status due to its context-dependent, i.e., previously established, discourse topics. The exceptions to this assumption are through rhematic progression, thematic breaks and the first sentence in the discourse, whose sentence constituents only have NEW information status. In rhematic progression, information status is switched such that the theme becomes NEW and the rheme becomes GIVEN. In thematic breaks, NEW themes are introduced either deliberately to indicate a rhetorical shift or unintentionally due to gaps in the text's logic. NEW themes are otherwise found in marked themes as context-independent discourse topics. If unmarked, then the marked theme has GIVEN information status. The present work does not divide rhemes into marked and unmarked rhemes.

Realizationally, all constituents up to and including the grammatical subject (and its dependents) constitute the overall theme. If present, the grammatical subject of the independent matrix clause becomes the **grammatical theme**. The term *assessment* from Figure 4-1 functions as the grammatical subject of the sentence without any fronted thematic elements and therefore becomes the grammatical theme.

UNMARKED THEME	RHEME
GRAMMATICAL THEME	**RHEME**
Assessment	*is a key component of any educational programme.*

Figure 4-1: Standard, unmarked theme structure without any fronted elements before the grammatical subject. The grammatical subject congruent with the finite verb in the independent matrix clause is defined as the grammatical theme.

The grammatical theme is the most basic thematic constituent a clause can possess and is therefore considered unmarked. Note that the terminological addition of *grammatical* to the base term *theme* is merely for delineation purposes: Fundamentally, the grammatical theme functions as the unmarked subject of the α-clause (independent matrix clause) or the unmarked subject of a projected β-clause (subordinate clause), as is the case in projecting themes (see Figure 4-7 below). This nomenclature conceptually and programmatically aids in distinguishing it from any marked themes that may be realized in the same sentence. If no theme is present or if a fronted circumstantial inverts word order, as in *Into the room came the man*, then there is no grammatical theme.

The next group of standard unmarked themes are found in interrogatives. In WH-interrogatives, standard or with a nominal phrase, the function of the interrogative is to inquire about missing information; it is the interlocutor's task to provide content that answers the question posed by the WH-interrogative (Halliday & Matthiessen 2014: 101). For polar interrogatives, the yes/no function of the question comes into the fore and is embodied through the finite verb.

Beginning with non-polar interrogatives, singular WH-interrogatives, like *how* in (1) in Figure 4-2, form the unmarked grammatical theme with the remainder constituting the rheme. If nominal phrases form part of the WH-interrogative as in (2) and (3), then the entirety of the

WH-interrogative phrase becomes the grammatical theme. For polar interrogatives, both the auxiliary verb and the congruent subject form as the grammatical theme; everything after the grammatical subject is then the rheme. This treatment is akin to WH-interrogatives with a nominal phrase, such that both the question initiator and the experiential topic are subsumed under the theme. This approach accounts for both the polar function inherent to polar interrogatives and the contribution that the grammatical subject affords to the thematic development of a sentence.

	UNMARKED THEME	RHEME
	GRAMMATICAL THEME	**RHEME**
WH-Interrogatives	*1. How*	*will that happen?*
WH-Interrogatives with Nominal Phrases	*2. Which train*	*did you want?*
	3. For whom	*did they explore the caves?*
Polar Interrogatives	*4. Did you*	*see that?*

Figure 4-2: Unmarked grammatical themes in polar and WH-interrogatives as employed in the present work.

Aside from the exception cases of clefts and existentials, this concludes the treatment of standard SVO sentences without any elements fronted before the grammatical subject of the independent α-clause. The remaining tables in the following address the use of marked themes, clefts, existentials and finally sentences without themes.

As mentioned in Chapter 4.1, marked themes are categorized into structural, modal, circumstantial, hypotactic and projecting themes. Regardless of whether sentences contain a marked theme, the grammatical subject of the independent α-clause (matrix clause) following immediately thereafter is always the unmarked grammatical theme where present. Again, this represents a deviation from traditional approaches in order to allow one unmarked theme (the grammatical subject in α-clauses or in a projected subordinate β-clause) and one or more marked themes in each sentence.[15] While marked themes support the contextualization of the ensuing text, the unmarked theme may facilitate the resulting thematic progression.

Figure 4-3 shows the thematic structure of a sentence containing both a marked **structural theme** and an unmarked grammatical theme. Since the sentence-initial *furthermore* is neither the grammatical subject nor does it function as point of departure in terms of the clause's communicative message, it is demarcated as a marked theme.

MARKED THEME	UNMARKED THEME	RHEME
STRUCTURAL THEME	**GRAMMATICAL THEME**	**RHEME**
Furthermore	*its external surface*	*consists of 12-ring cups.*

Figure 4-3: Structural themes fall under fronted adverbials that function as a cohesive, signposting device between sentences.

Structural themes as found in Figure 4-3 act as a cohesive device to establish logical links between sentences. Their employment at the beginning of a clause merits their marked nature since they guide the reader in a specific logical direction on the basis of the previous clause and the information to come after the structural theme. Due to their sign-posting nature, structural themes form a closed class of adverbials, such as *in addition*, *because of*, *contrarily*, *nonetheless* and many more (cf. Huddleston et al. 2021: 208-228).

[15] There may only be one unmarked theme in a sentence, but that, in turn, may be comprised of multiple sentential elements, e.g. *The boy and girl left the house*, where *The boy and girl* is the unmarked theme consisting of a coordinated noun phrase.

MARKED THEME	UNMARKED THEME	RHEME
MODAL THEME	**GRAMMATICAL THEME**	**RHEME**
Specifically,	*we*	*used ADNI1 baseline dataset for our model.*

Figure 4-4: Modal themes are a class of fronted adverbials that insert (inter)personal, subjective, specifying or evaluative information into the discourse message.

Modal themes as in Figure 4-4 are also inherently marked due to their function of construing information to be interpreted in a personal, evaluative, specifying or subjective light (Ma & Zhu 2023: 467-468). While subjective modal adverbials may appear less in formal, academic environments, their appearance is commonplace in discourse with greater interpersonal and interactional content, such as novels, lyrics or blogs. Modal adverbials enjoy a wider range of lexical expression, and their identification was oriented around Halliday & Matthiessen (2014: 108-109).[16]

MARKED THEME	UNMARKED THEME	RHEME
CIRCUMSTANTIAL THEME	**GRAMMATICAL THEME**	**RHEME**
According to Kline,	*a sample size for SEM*	*should be more than 100.*

Figure 4-5: Circumstantial themes as fronted prepositional or temporal noun phrases that can indicate location, temporality, manner, contingency, matter, cause, angle or accompaniment.

The next class of marked themes, **circumstantial themes**, is given in Figure 4-5. The example there represents a deviation from the Hallidayan approach in that circumstantial adjuncts, here *according to Kline*, do not constitute the sole (topical) theme. The Hallidayan approach would identify the marked theme as *according to Kline* and everything thereafter the rheme – including the grammatical subject *a sample size for SEM*. This can affect which discourse topics are developed as the foundation of the message (theme) and which as the core of the message (rheme). In the present work, however, circumstantial adjuncts are defined as a standalone class of thematic adjuncts called circumstantial themes. Through standalone treatment, both marked circumstantial themes and the accompanying unmarked grammatical theme can be readily traced.

Circumstantial themes represent the most complex theme type in that a myriad of prepositional and temporal noun phrases can instantiate their realization. Additional examples of circumstantial themes are *in the future*, *yesterday*, *two years ago*, *of late*, *on the 2nd page*, *under the table*, *from whence it came*, *to some* and *on these grounds*, to name a few. As these examples illustrate, circumstantial themes can demonstrate TEMPORAL, CAUSATIVE, LOCATIVE, ANGLE, MANNER, CONTINGENCY and even ACCOMPANIMENT (*along with*) relations. While structural themes facilitate logical and structural development between sentences via cohesive devices, circumstantial themes establish the semantic stage for the information that follows. Their use thereby epitomizes the contextualizing function that fronted adjuncts may have.

Next is the class of **hypotactic themes**, whose denotation stems from the hypotaxis they reflect, as reflected in Figure 4-6. In this case, the hypotactic β-clause *If they agree to participate* becomes the hypotactic theme. This theme type must fulfill two conditions in order to be called such: first, a subordinating adverbial must introduce the β-clause. Examples of hypotactic adverbials that can instantiate a hypotactic theme are *because, since, if, when* and *once*. Second, the β-clause must have its own grammatical subject and congruent finite verb. Non-finite relative clauses, such as *Having drunk the coffee, the woman left the café*, form an exception

[16] cf. also Frey (2003).

since the grammatical subject is elided in the hypotactic clause but realized in the matrix α-clause. Again, according to the Hallidayan approach, hypotactic themes alone would have constituted the topical theme of the sentence.[17] Extending the overall theme's scope to include both the hypotactic clause and the grammatical subject in the matrix α-clause facilitates pinpointing how the theme is developed across sentences with hypotaxis.

Subordinate β-clause	Matrix α-clause	
MARKED THEME	UNMARKED THEME	RHEME
HYPOTACTIC THEME	**GRAMMATICAL THEME**	**RHEME**
If they agreed to participate,	*the questionnaire*	*would be distributed.*

Figure 4-6: Hypotactic themes as subordinate adverbial clauses that appear sentence initially. Their classification is merited through the clause-initial adverbial, syntactic dependence as a clausal complement of the matrix clause's finite verb and a subject with congruent finite verb within the subordinate clause.

Similar to circumstantial themes, hypotactic themes aid in contextualizing the rhematic information that follows. Further, depending on the subordinating adverbial used, CONDITION, TEMPORALITY, CONCESSION and MANNER relations can be expressed. The semantic relations afforded by hypotactic themes are noticeably less. Yet, hypotactic themes are able to provide further specification or qualification of the propositional content on account of their inclusion of a subject and finite verb within the dependent clause. This allows for the introduction of NEW discourse topics embedded along GIVEN ones either directly within the hypotactic theme or the following grammatical theme.

The final class of marked themes is **projecting themes**, which is exemplified in Figure 4-7. There, three forms of projecting themes are illustrated that all follow the same thematic structure. Regardless of projection, cleft or thematic equative, the matrix α-clause functions as a projecting clause complemented by a subordinate and projected β *that*-clause. The matrix α-clause becomes the projecting theme; the subordinate clause after the *that*-adverbial then has its own grammatical theme and rheme.

The categorical distinction between **projections**, **clefts** and **thematic equatives** stems from the syntactic and lexicogrammatical construction of the α-clause. In projections, the verbal clause in the matrix sentence denotes a mental, material or relational process characterized by the pragmatic function of 'saying' per Halliday & Matthiessen: these 'saying' verbal clauses can establish dialogue in narrative and reporting, inform sources in newspapers, or even quote and paraphrase in academic contexts (cf. Halliday & Matthiessen 2014: 302–305).

Clefts as shown in Figure 4-7 employ a non-referential dummy-*it* and a copular predicate, followed by the subordinate *that*-clause. For a cleft to be categorized as a projecting theme, it must follow this structure as other forms of clefts exist. For example, *It's not necessary to write your name* is another type of cleft that concludes with an infinitive clause. This class of clefts is treated exceptionally in the present work's thematic analysis and described below. Functionally, clefts are predominately used as a means for specification whose marked structure is revealed through a shift of GIVEN information to the subordinate *that*-clause (Patten 2012: 3-

[17] It should be noted that Halliday does indeed argue for splitting subordinating clauses into additional thematic and rhematic constituents (Halliday & Matthiessen 2014: 125-127). However, he argues that either a bipartite, i.e., one theme and one rheme only, or multipartite analysis should be employed. In the present work, the marked theme is analyzed in conjunction with the unmarked grammatical theme and rheme. Complex marked themes, such as hypotactic and projecting themes with their own subordinate, dependent structure, are not broken down into their own theme-rheme structure, however.

4). This marked structure is reinforced through the use of the non-referential dummy-*it* which serves as a rhetorical marker.

	Matrix α-clause	Subordinate β-clause	
	MARKED THEME	UNMARKED THEME	RHEME
	PROJECTING THEME	**GRAMMATICAL THEME**	**RHEME**
Projection	*The results obtained indicated that*	*the response*	*was the strongest.*
Cleft	*It is important that*	*the remainder*	*can be tested.*
Thematic Equative	*What is for sure is that*	*the information*	*is lacking.*

Figure 4-7: Projecting themes shift the GIVEN discourse topic to after the matrix α-clause and within the subordinate, projected β-clause. The subordinating adverbial (that) *marks the end of the projecting theme and is one of the requirements for a projecting clause. The second requirement is a cleft structure, a thematic equative or projecting verb.*

Thematic equatives form the final form of projecting theme and are based on the self-same term put forth by Halliday & Matthiessen (2014). They define thematic equatives as a bi-constituent clause whose copula serves as an equating pivot between both clauses (Halliday & Matthiessen 2014: 93-94). In other words, in *What I need is help*, the thematic constituent *What I need* is equated to the rhematic constituent *is help*. Hence, an equative "X is Y" relationship is formed through clausal elements X and Y. Where the present work differs from this approach is when a subordinate *that*-clause follows the copula, as in Figure 4-7. Here, all constituents up to and including the *that*-adverbial constitute the projecting theme, similar to projections and clefts. The grammatical subject of the subordinate *that*-clause then becomes the grammatical theme, and the remainder becomes the rheme. This approach allows for a uniform treatment of such marked structures despite their varied realizational patterns and rhetorical functions.

It could be argued that the sentences in Figure 4-7 should be analyzed as having a standard thematic structure: *The results obtained* as the unmarked theme and the remainder of the sentence the rheme, in the case of projection, for example. However, understanding the effect such projecting themes have on the information status of the sentence constituents can explain the present work's approach to their thematic analysis. Projecting clauses shift the GIVEN discourse topic to the subordinate *that*-clause, with the grammatical subject of the matrix α-clause most commonly being NEW. Hence, the foundation of the discourse message (*the response* in Figure 4-7) is sandwiched between the NEW projecting α-clause and the NEW rheme that follows. The projecting theme becomes marked through its rhetorical or process function (as per Halliday & Matthiessen 2014), such that it informs, reports, projects or specifies the GIVEN discourse topic. This discourse foundation is finally pushed forward by the rheme that concludes the β *that*-clause and sentence on the whole.

The present treatment of projecting themes can be further substantiated with the following: the projecting clause functions as a cataphoric marker of the foundation of the discourse message to come in the β-clause, i.e., the grammatical theme. The projecting theme thereby creates tension that must be resolved through the given grammatical subject and rheme within the *that*-clause. Treating projecting clauses as such allows combining both the thematic structure and its rhetorical function into a single construct. Isolating such structures as standalone marked themes therefore helps to highlight such rhetorical functions in writing.

The penultimate class of thematic structures to consider begins the present work's treatment of exception cases as first outlined in Chapter 2.6. Figure 4-8 exemplifies **non-projecting clefts** and **existentials**, whose analysis is inspired by but again deviates from the Hallidayan approach.

The cleft example falls under an adjectival cleft, whereby the dummy-*it* has no anaphoric co-referent but acts cataphorically to the accompanying *adj* + *infinitive* structure. The content in the infinitive clause thereby resolves the tension established by the cleft *it is*, similar to how the propositional content within the *that*-clause resolves the tension from the projecting theme. To embody the cleft structure and its rhetorical function of expressing an angle, opinion or stance, *it is* is qualified as the grammatical theme and the remaining *adj* + *infinitive* clause as the rheme.

	MARKED THEME	RHEME
	GRAMMATICAL THEME	**RHEME**
Cleft	*It is*	*important to understand the pulse of money market and capital market.*
Existential	*There is*	*little supporting scientific evidence.*

*Figure 4-8: The two exception cases of non-projecting clefts and existentials are denoted as grammatical themes through the dummy-*it *and copula for clefts and the existential* there *and copula for existentials. Their exceptional treatment is merited through their non-referential, rhetorical function in discourse.*

Following the same practice of encapsulating rhetorical function in thematic structure, the existential adverbial *there* together with the copula or copula complex (e.g., *may have been* as in *there may have been*) is considered the grammatical theme. All constituents after the existential structure then belong to the rheme. Rhetorically, existentials draw the reader's attention to the propositional content in presentational, NEW form. They place emphasis on the introduction of NEW participants, circumstances, conditions or objects within the discourse and underline the rheme's NEW status through their presentational function (Halliday & Matthiessen 2014: 308). While some argue in favor of including the experiential element of an existential (i.e., the noun phrase following the copula) into the theme (cf. Davies 1997), the danger in doing so would be the potential for a text without a rheme. As rhemes are considered obligatory in thematic analyses, in contrast to themes, extending the theme span to the entire existential phrase would violate this theoretical condition. Further, qualifying the NEW experiential element as the theme would contradict its function as the GIVEN foundation of the discourse message. Finally, since the *there* is neither deictic nor coreferential, as is the case with a dummy-*it*, its thematic progression can be considered either a thematic break or rhematic progression. As outlined below, the present work opts for the latter.

The final case to consider in determining the thematic status of sentence constituents is that of **imperatives** and **fragments** as shown in Figure 4-9. With imperatives, the subject functioning as the recipient of the message can either be elided or realized as a vocative. In either case, however, the assumed *you* (as in *You go!* or *You wait, George*) remains unrealized, which results in the finite verb alone constituting the bare minimum element of an imperative. Since there is no explicit grammatical subject, the entire imperative becomes the rheme instead of relegating it to the theme. In fragments, the text is realized as an incomplete, dependent clause. This may be due to the absence of a subject-verb pair or the realization of syntactically dependent phrases alone, such as prepositional phrases. Should a text be realized as an adverbial phrase or dependent clause, therefore, they are considered a fragment and entirely rhematic. Analytically, these then most commonly cause a thematic break or possible rhematic progression depending on whether connecting thematic or rhematic constituents can be identified in the previous sentence.

	THEME	RHEME
	GRAMMATICAL THEME	RHEME
Imperative	-	*Go!* *Wait, George!*
Fragment	-	*Because of the time they needed.* *Reading the book.*

Figure 4-9: Imperatives and fragments as entirely rhematic structures constituting an exception class in their thematic analysis.

In natural language, texts commonly exhibit more complex thematic realization patterns than the simplified examples illustrate in the previous tables. An individual sentence can often possess multiple marked themes and exceptional structures together, which are individually analyzed but collectively constitute a sentence's overall thematic structure. For example, in formal texts, sentences containing a structural, projecting, circumstantial and grammatical theme are not rare, e.g., *However, it is argued that, under certain conditions, the former case fails.* Here, only *fails* would constitute the rheme with all other sentence constituents falling into corresponding theme classes. Thematizer's functionality behind dissecting such complex formulations into their individual thematic and rhematic constituents rests upon the theoretical framework outlined in this section.

4.4 Thematic Progression Patterns Employed in the Present Work

The theme and rheme definitions from Chapter 4.3 underlie the thematic progression patterns that are used in the present work and discussed in the following. The examples of each thematic progression pattern provided in the following stem from the sample texts used for developing and training Thematizer. A brief explanation of how the thematic progression pattern is instantiated via thematic and rhematic constituents then follows. The effect that each thematic progression pattern has on text development complements their treatment. This collective discussion represents the conceptual and theoretical framework behind Thematizer's identification and analysis of thematic progression in text.

In **constant continuous progression**, the same theme is realized across two concomitant sentences. Here, the pattern $T_1 \rightarrow T_2$ ensues, whereby the same GIVEN discourse topic is realized thematically (cf. Figure 4-10). Through this realizational pattern, the same foundation of the discourse message is developed across two sentences. In doing so, greater salience emerges in the propositional content from the NEW discourse topics within the rheme as a means to push forward discourse. Constant continuous progression is adopted from the selfsame model originally proposed by Daneš (1974) and belongs to the basic inventory of progression patterns employed in thematic analyses.

CONSTANT CONTINUOUS PROGRESSION

	THEME	RHEME
Sentence 1	***We*** [T_1]	*were involved in all the steps of calculating proportions for the mix itself.*
Sentence 2	***We*** [T_2]	*also used testing procedures such as the slump test, unit weight, and amount of air entrained per unit volume.*

*Figure 4-10: Constant continuous progression occurs when the theme of the first sentence [T_1] is instantiated as the theme of the second sentence [T_2], resulting in a [T_1] → [T_2] structure. Pertinent thematic elements that contribute to the progression pattern are highlighted in **bold**.*

The second form of progression belonging to the basic inventory of thematic progression from Daneš is **simple linear progression** and shown in Figure 4-11. In this pattern, a rhematic element from one sentence is instantiated thematically in the subsequent sentence, such that $R_1 \rightarrow T_2$. Here, a NEW discourse topic from the first sentence is instantiated as the GIVEN foundation of the message in the subsequent sentence. This zig-zag pattern is common in more formal text types and affords greater dynamicism between sentences through re-instantiation of previously NEW information as GIVEN (Rosa 2013: 221). As shown through the paraphrase in Figure 4-11, the entire rheme can be instantiated as the theme in the subsequent sentence.

SIMPLE LINEAR PROGRESSION

	THEME	RHEME
Sentence 1	*These conditions*	***include what type of structure is being formed, the thickness of the forms, reinforcement requirements, consolidation methods, and most importantly the compressive strength needed.*** [R_1]
Sentence 2	*Once* ***this array of input*** [T_2] *is acquired, the mix proportioning process*	*can begin.*

Figure 4-11: Simple linear progression re-instantiates rhematic element(s) from the first sentence [R_1] as the theme in the second sentence [T_2], resulting in a [R_1] → [T_2] structure. Pertinent rhematic and thematic elements that contribute to the progression pattern are highlighted in ***bold****.*

If intermediary sentences are inserted between constant continuous or simple linear progression, then the corresponding gapped progression pattern is present. This is based on split-theme progression as originally postulated by Dubois (1987), Rørvik (2012) and Hawes (2015). Whereas constant continuous progression occurs across the themes of two concomitant sentences, **gapped constant progression** manifests when the theme is re-instantiated two or, at most, three sentences later (cf. Figure 4-12). In doing so, the progression pattern $T_1 \rightarrow T_3$ (if realized thematically two sentences later) or $T_1 \rightarrow T_4$ (thematic realization three sentences later) emerge.

GAPPED CONTINUOUS PROGRESSION

	THEME	RHEME
Sentence 1	*The amount of coarse* ***aggregate*** [T_1] *to be used in the mix*	*must now be determined.*
Sentence 2	*This*	*can be done utilizing Table 10.8.*
Sentence 3	*Based on the maximum aggregate size and the fineness modulus of the* ***fine aggregate*** [T_3]*, a value*	*can be found in the table.*

Figure 4-12: Gapped continuous progression indicates the development of a theme [T_1] two or three sentences after its initial mention, resulting in a [T_1] → [$T_{3/4}$] structure. Intermediary progression occurs until the initially thematic element is realized as a theme at a maximum of three sentences later. Pertinent thematic elements that contribute to the progression pattern are highlighted in ***bold****.*

Conversely, if the rheme of one sentence is developed as the theme two or, at most, three sentences later, then **gapped linear progression** is at hand. As Figure 4-13 illustrates, the rhematic element *older plants* is paraphrased and realized as the theme in the third sentence, resulting in a progression of $R_1 \rightarrow T_3$. While gapped progression (constant or linear) is interrupted by intermediary sentences, that does not mean that progression ceases between the elements that initiate the gapped progression. In Figure 4-13, the progression pattern between the first and second is simple linear due to the meronymous relationship between the rhematic *older plants* and thematic *branchlets*. The text then returns to the rhematic element *older plants* from the first sentence and instantiates gapped linear progression through the thematic *the mature trees*.

GAPPED LINEAR PROGRESSION

	THEME	RHEME
Sentence 1	*The bark*	*is generally dark brown to grey—smooth in younger plants though it can be furrowed and rough in* ***older plants*** [R_1].
Sentence 2	*Branchlets*	*may be bare and smooth or covered with a white bloom.*
Sentence 3	***The mature trees*** [T_3]	*do not have true leaves but have phyllodes—flat and widened leaf stems—that hang down from the branches.*

Figure 4-13: Gapped linear progression is a complementary pattern to simple linear progression, whereby the rheme of the first sentence [R_1] is realized as the theme two to three sentences later, resulting in a [R_1] → [$T_{3/4}$] structure. Intermediary progression occurs until the initial rhematic element is realized as a theme at a maximum of three sentences later. Pertinent rhematic and thematic elements that contribute to the progression pattern are highlighted in ***bold****.*

Gapped progression patterns are most commonly employed in scientific and formal texts, where discourse topics are enumerated and treated individually (Dubois 1987: 95). Their complex patterning reflects the comparatively complex subject matter being explained, although content complexity is not a requirement for use of this pattern. Finally, while no upper limit has been formally defined as to how far back gapped progression should occur, the present work follows the suggested approach of limiting gapped progression to two to three sentences prior (McCabe 1999, Jalilifar 2009). The further the reader must refer back to previously established topics, the greater the risk of compromising comprehensibility. For that reason, a shorter sentence range was selected to uphold comprehension and, as will be outlined in Chapter 4.10, to limit the computational complexity in determining thematic referents across multiple sentence ranges.

Rhematic progression is an additional pattern that Daneš did not originally postulate and is derived from the work by Enkvist (1973), Cloran (1995), Shi (2013) and Dou & Zhao (2018). Here, the rheme remains the center of focus in discourse development, which makes the thematic discourse topic more salient. Rhematic progression either manifests as shown in Figure 4-14, whereby $R_1 \rightarrow R_2$, or when the theme is instantiated as the rheme in the subsequent sentence, $T_1 \rightarrow R_2$. In the former case, the NEW discourse topic presented in the first sentence is exceptionally developed as a GIVEN rhematic element in the next sentence. Conversely, in a $T_1 \rightarrow R_2$ progression, the GIVEN theme in the first sentence is realized as a GIVEN rheme is the second. In either case, the theme in the second sentence becomes marked and NEW since the rheme is afforded GIVEN informational status. These are rather rare, marked structures since the conventional GIVEN → NEW information structure is flipped. Their salient use therefore concentrates around drawing attention to a NEW foundation of the discourse message as a shift to a new rhetorical section in the text.

RHEMATIC PROGRESSION

	THEME	RHEME
Sentence 1	*Many times this*	*is* ***limited to*** [R_1] *the aggregate that is offered in the area or a preference to what should be used in the concrete.*
Sentence 2	*For design with no supply restrictions, there are*	***three limitations*** [R_2] *to take into account.*

Figure 4-14: Rhematic progression takes place across two concomitant rhemes, such that a NEW discourse topic [R_1] is developed as GIVEN in the subsequent rheme [R_2], resulting in a [R_1] → [R_2] structure. Alternatively, the theme of the first sentence [T_1] can be developed as the rheme of the second sentence [R_2], i.e., [T_1] → [R_2] (this progression is not shown in the figure). Pertinent rhematic elements that contribute to the progression pattern are highlighted in ***bold****.*

Thematizer automatically parses non-projecting clefts with a dummy-*it* and existentials as rhematic progression. The reason for this, as explained in Chapter 4.3, is due to the rhetorical function these exceptional structures hold. Both clefts and existentials establish tension due to

their lacking coreferentiality and cataphoric nature. Once the discourse topics are realized after the cleft or existential, the tension is resolved and cataphoric function fulfilled. This further embodies the rheme's function of pushing the discourse message forward. In order to accentuate these structures' exceptional use case and function, they are automatically assigned rhematic progression.

The next class of thematic progression included in the present work's theoretical model is that of **macrothemes**. These are based on their original conceptualization and treatment in Martin (1992) and based on Daneš's hypertheme. Macrothemes can be understood as statistically significant discourse topics at the macro level of the text. In other words, these represent the primary topic(s) around which the entire text revolves, and from which related discourse topics treated at the paragraph or document level are derived. In the example provided in Figure 4-15, the text discusses the topic of concrete, which is introduced as a GIVEN discourse topic in the first theme. Derivatives of this topic are evident in the subsequent sentences through the rhematic meronymous realizations *coarse aggregate* and *admixtures*. These account for rhematic progression in the intermediary sentences until the macrotheme *concrete* is re-instantiated in sentence six.

MACROTHEME INSTANTIATION

	THEME	RHEME
Sentence 1	***Concrete*** [T_1]	*is made through appropriate admixtures.*
Sentence 2	*The basic procedure*	*begins by adding all coarse aggregate into the mixer with about half the water.*
Sentence 3	-	*Next add the fine aggregate, all cement, and the remaining water.*
Sentence 4	-	*Finally add the appropriate admixtures and allow it to mix for 3 minutes.*
Sentence 5	-	*Let stand for 5 minutes then continue to mix for 3 more minutes.*
Sentence 6	***The concrete*** [T_6]	*is then thoroughly mixed and ready for testing and use.*

*Figure 4-15: Macrothemes, as indicated in **bold**, are statistically significant discourse topics that reflect the overarching discourse at the paragraph and document level. They are determined computationally via Latent Dirichlet Allocation and are instantiated so long as a minimum of three sentences exists between their last realization.*

In order for macrotheme instantiation, i.e., progression, to be present, two conditions must be fulfilled: firstly, the thematic element to be realized must be considered a statistically relevant discourse topic. This is determined through Latent Dirichlet Allocation during thematic progression classification (see Chapter 4.10.4). Secondly, macrotheme instantiation can only be realized if the macrotheme is not present in the previous three sentences. Since *concrete* is only employed thematically in the first and sixth sentence, a minimum gap of three sentences merits its re-instantiation as the macrotheme. Had *concrete* been used at any point in sentences two to five, then constant, linear or gapped progression would have been present on account of lexical repetition. Therefore, if standard or gapped progression can account for the development of thematic or rhematic elements within a span of three previous sentences, macrotheme instantiation cannot ensue. This not only ensures that macrothemes are not overly abundant and redundant, but also ensures that the other progression patterns are first used to determine the potential means of progression.

The final form of thematic progression for the present theoretical model is the deliberate or unintentional lack of progression across sentences via **thematic breaks**. These are present when none of the previously elaborated thematic progression patterns can be identified. In other words, no themes or rhemes from the previous three sentences (or entire text for macrothemes) develop the discourse across sentence clusters. This commonly occurs when a sentence is

realized without a theme, as shown in the example from Figure 4-16. However, since rhematic progression can also account for progression in expressions without a theme, sentences containing rhemes alone do not automatically equate to a thematic break.

THEMATIC BREAK

	THEME	RHEME
Sentence 1	*The assumptions made*	*are listed below.*
Sentence 2	-	*Working Hours.*

Figure 4-16: Thematic breaks occur when neither thematic nor rhematic elements are developed across concomitant sentences. While most common between rhetorical sections of a text, they can also be used stylistically to raise attention to a particular passage in text.

Thematic breaks are often employed to indicate a new section in the rhetoric, such as transitioning to a new subsection or chapter in a book or highlighting particularly important information in the argumentative development of a text. That being said, inappropriate or frequent thematic breaks could be indicative of poor logical development in a writer's text. Further, inappropriately using a NEW discourse topic thematically can also initiate and substantiate a thematic break. It is with these final two points where an overabundance of thematic breaks can serve as insightful input to the logic and structure of writers' text.

Together with the formal definition of marked themes, the grammatical theme and rheme outlined in Chapter 4.3, these thematic progression patterns form the theoretical foundation of the present work. While much inspiration was taken from the theoretical framework established by Halliday and the Prague School, key additions to the thematic progression patterns reflect contemporary expansions to Daneš's originally proposed models. Taken together, these models informed the theoretical basis of Thematizer's development and functionality.

4.5 Training and Test Materials for Thematizer's Development

In this section, the texts used for training Thematizer during development are introduced. Furthermore, the reasoning behind the specific text types chosen for training are outlined with a summary of the texts' qualitative parameters. Finally, the section concludes with the timeframe needed for training and development.

In order to develop and test the performance of Thematizer, five different text types were used: English-native (L1) university texts, English-non-native (L2) university texts, Wikipedia articles, blog posts and lyrics. Both groups of university texts were written tasks completed at the undergraduate or graduate level. Their text types covered term papers, scientific reports, creative writing pieces, argumentative essays, abstracts, proposals, response papers, critiques and formal letters. The texts themselves stemmed from the humanities and sciences. Whereas the L1 university texts were taken from the Michigan Corpus of Upper-Level Student Papers (2009), all L2 university texts were given as tasks in English for Specific Purposes courses at Friedrich-Alexander University, in Erlangen, German.[18] The students of these courses were predominantly German L1 speakers. Non-native samples of writing were included to determine whether the thematic structure in their texts differed from the L1 university texts as well as the other text types. Additionally, particularly problematic syntactic patterns present in L1 texts were used to train their recognition during thematic analyses, since these otherwise may have caused Thematizer to crash or impair overall functionality during parsing.

[18] All texts were made anonymous and no mention of the original author is given or available for data privacy reasons.

Wikipedia articles were chosen due their ease of access and more formal register compared to blogs and lyrics. Due to their expository and textbook-like nature, Wikipedia articles form a complementary text type to L1 university texts that reflect similar texture characteristics. Wikipedia articles also form a somewhat unique test case in that they were written by at least one author. Multiple authorship for Wikipedia articles may have affected how the texts were developed thematically, which Thematizer could potentially identify in its analyses.

Blog posts were also chosen for their ease of access but also as a representation of written, evaluative, expository, commercial and/or informal text. Blog posts were chosen at random from Google search results on 'popular blogs' while collating other texts for analysis in May 2022. Naturally, algorithmic bias as to what constitutes 'popular' and as to how Google parses search results through keyword inflation on websites should not be ignored. Blog post texts used for training ranged from recipes and travel blogs to game reviews and self-help articles.

Finally, lyrics were used as a juxtapositional text to the more formal text types, particularly since lyrics are fundamentally creative and poetic in nature. While lyrics are meant to be sung, their poetic and prepared (i.e., not spontaneous) characterization merits their analysis from a written account. Their degree of informality compared to the other text types was a further determining factor in their selection. This informality is reinforced through the frequency of dependent and/or incomplete structures (e.g., vocatives, interjections, elided subjects in independent clauses and standalone gerunds), which would conventionally be considered ungrammatical in texts with formal register. Lyrics were then used to test Thematizer's analytical and parsing functionality when confronted with texts that follow less stringent (i.e., less conventional) syntactic realization requirements. Finally, lyrics were chosen because little research has been done on this text type in terms of its thematic progression distribution (cf. Pagih 2019).

Altogether, 30 texts were used for each text type. This resulted in a total of 150 texts employed for the development and training of Thematizer. The average sentence and token length of each text, including standard deviation, can be found in Table 4-3. Once retrieved from their respective source, texts were converted to a .txt file to be processed by Thematizer. No corrections were made to grammatical or clerical errors evident in the texts. The source for each text can be found in the primary sources in the bibliography at the end of the dissertation.

Text Type	Sentence Length	Token Length
Wikipedia Articles	67 (± 10)	1557 (± 165)
L1 University Texts	67 (± 27)	1414 (± 421)
L2 University Texts	23 (± 15)	477 (± 301)
Blog Articles	70 (± 36)	1010 (± 388)
Lyrics	41 (± 22)	303 (± 123)

Table 4-3: Average length of the five text types used for the training of Thematizer, given in terms of sentence and token length and with corresponding standard deviations in parentheses.

As the texts were either sourced from readily available online resources or anonymized repositories, the text collation process was completed over the course of a single week in May 2022. Development of Thematizer began in September of 2021 before training texts were

required for more wide-scale testing. Development continued until September 2022 with the use of the entire training text set, at which point the final version of Thematizer was completed.

Testing the software with the training texts was conducted parallel to the developmental process. This involved the comparison of Thematizer's automated output to the manual analyses of all 150 texts' thematic structure that the present author conducted beforehand. Errors from the automated analyses were collated and categorized according to either functional (i.e., programming-induced) or analytical (i.e., thematic misparse) errors. These errors served as an indication of Thematizer's performative results and parsing accuracy. Once changes were made to the code, all texts were passed through Thematizer again for another round of comparison between the manual analysis and automated output. Re-analysis of the automated output compared to the manual analyses took approximately between three weeks and one month for each version update. Altogether, Thematizer required seven version updates to account for the various error types and to ensure the program's foundational functionality with the help of the training texts.

Finally, for validation purposes, ten novel texts that Thematizer had not been trained on were randomly selected to test its parsing functionality and accuracy. The ten text types selected were an article from a gaming news site, a newspaper article, an excerpt from a linguistics textbook, Reddit comments, an editorial, an obituary, the comments section from a blog entry, a Wikipedia article, an L1 university text and a short story. On average, each text had approximately 1,232 (± 107) tokens and a length of 67.4 (± 19.6) sentences. These text types were chosen to reflect a range of register and texture characteristics of written language, particularly with text samples that Thematizer had had no previous exposure to. The inclusion of a novel L1 university text and a Wikipedia article was to determine whether similarities in parsing accuracy occurred between training and test texts. No changes were made to the text before parsing. The source for each text can be found in the primary sources in the bibliography at the end of the dissertation.

4.6 Program Specifications behind Thematizer

Thematizer was written predominantly in the programming language **Python**. This language was primarily chosen because it is available in all operating systems and most development environments. The number of application programming interfaces and libraries developed for use in Python-driven natural language processing applications lends credence to the reliability of Python for language-specific purposes (Sarkar 2019). Second, while Java and JavaScript are also standard languages for natural language processing (see Stanford CoreNLP), Python has seen an expansion in its use particularly for natural language processing. This is largely due to the increased computing performance and power of standard hardware (Thelin 2021). Third, application deployment has become far more streamlined across all programming languages, but Python again has proven to be the most robust in terms of convenience and program complexity when shipping an application from development to production (Sanner 1999).

Finally, and perhaps most importantly, Python was chosen for the present work so as to use the application programming interface **Spacy** (Honnibal et al. 2020a). Spacy was written in Python for Python users and conceptualized specifically for natural language processing applications. It was developed as an alternative to Stanford NLP, Natural Language Toolkit (NLTK), Stanza, OpenNLP and other natural language processing toolkits. It implements machine-learning-based transformers for state-of-the art accuracy rates in terms of part-of-speech tagging, dependency parsing and named entity recognition. Spacy's third and current version has a 92.0% parsing and 97.4% tagging accuracy rating at processing speeds 10 times faster than other

applications (approximately 10,000 words per second (wps) compared to the average 1,000 wps from its competitors Stanza, Flair and UDPipe, see Honnibal et al. 2020d). Thematizer employs several of the natural language processing tools that Spacy offers for the syntactic and thematic analyses, more specifically part-of-speech tagging, dependency parsing, named entity recognition and lexical entailment.

In order to account for coreference resolution in text, the library **Coreferee** (Hudson 2022) was used, which is of particular importance for thematic progression analysis. This Python 3 library can be readily integrated into Spacy programs and is based on both neural networks and static rules. For instances across sentences where neither lexical repetition nor macrothemes are present, coreference chains may account for the thematic progression at hand. Such chains are established in text through the realization of an antecedent (i.e., a noun phrase) and its coreferencing proform. For example, in the previous sentence, the possessive pronoun *its* establishes a coreference chain with *antecedent*. These chains can occur within and across multiple sentences. In turn, they may contribute to a text's overall method of development and therefore need to be tracked.

It should be noted that another coreference resolution tool, NeuralCoref, has already been developed specifically for Spacy. However, it requires versions of Spacy up to but not including 3.0. For Thematizer, word vectors and pre-trained models used for part-of-speech tagging and dependency parsing require Spacy 3.0+. Models without pre-trained word embeddings exist but require supplying one's own parsed data, which would have been beyond the scope of this work. That is why Coreferee was ultimately chosen as the tool for coreference resolution in this work.[19]

Another library used for thematic progression analysis was **Gensim (**Řehůřek & Sojka 2010, 2011). This was chosen for its ability to perform Latent Dirichlet Allocation (LDA), a generative statistical model that can be used for topic modeling. Latent Dirichlet Allocation specifically uses probability distributions by means of Bayesian inferences to provide statistically significant terms or co-occurring noun phrases in a text. These statistically significant topics are used as potential macrothemes that are instantiated thematically throughout the text. The (re-)introduction of macrothemes allows treating an established topic in light of previous discourse topics, i.e., previous thematic and rhematic elements within the text. Macrothemes account for lexical repetition that fails across concomitant sentences but succeeds across spans greater than three sentences prior. In order for repeated lexis to become a macrotheme, however, Gensim must have identified it as statistically significant within the textual context. How Gensim accomplishes this is outlined in Chapter 4.10.4 below.

These libraries and the Python framework then constitute the backend of the program. Since the frontend as the user interface was a critical component in the development of Thematizer, a web interface was chosen to permit design without any dependency on a specific operating system, installation requirements or application prerequisites.

The so-called **Dash** framework (Plotly Technologies Inc. 2015) was used to create the web interface for user input and interaction. In order to render web interfaces, Dash uses Plotly libraries that populate dynamic plots (bar, line, graph, etc.) from user input. Thematizer was embedded as the backend into the web interface frontend. It is therefore via Dash itself that the

[19] It appears that NeuralCoref is also no longer being actively developed ever since Spacy 3.0 was released. Currently, work is being done to develop an alternative to NeuralCoref that functions within and for Spacy 3.0+. For more information concerning the most recent developments, see:
https://github.com/explosion/spaCy/pull/7264#issuecomment-986957673.

text to be analyzed is fed into Thematizer as the parser. The data and results from Thematizer's thematic analyses are then returned to Dash, which presents the visualized data as interactable graphs, drop-downs, figures, tables, text and plots.
The data used to populate the input and results for output in the web interface is the cumulative result of all syntactic and semantic parses explained in Chapters 4.8 to 4.10. The results are saved in a JSON dictionary which contains each sentence's dependency parse, grammatical theme, rheme, marked theme, marked theme type, marked theme semantic class, thematic progression pattern, means of progression, connecting element that instantiates thematic progression and filename of the text entered or uploaded. This dictionary is passed on to the respective methods within *dashapp.py*, which is a separate Python file where the functionality of Dash resides.

As for the layout and design of the website interface itself, which Dash functionally supports but does not offer as a default, HTML and CSS were used. Where possible, settings were used that did not necessitate a specific browser. The functionality of the website itself via the interface was tested to work on Chrome, Safari, and Firefox, as these are the most commonly used browsers. As is the case with nearly all web applications, minor stylistic, but no functional, differences emerged across browsers.

In terms of the versions of the libraries, APIs and frameworks specifically, Thematizer was written with Python 3.8.0, Spacy 3.1.0, Coreferee 1.0.1, Gensim 4.1.2 and Dash 2.0.0. For text cleaning, data processing and data presentation with Dash, the libraries Dash Bootstrap Components 1.0.3, NumPy 1.19.5, Pandas 1.3.4 and Roman 3.3 were used. As this program and project will continue to be updated to address residual backend and frontend issues, updates to more recent versions of the APIs employed will ensue where possible and appropriate.

4.7 Thematizer Functionality and Parsing Tasks

On the basis of the programming environment, libraries and frameworks outlined in the previous section, the fundamental functionality behind Thematizer in terms of its programming will now be explained. The general processing steps that Thematizer progresses through for each text is shown summarily in Figure 4-17.

At the outset of the parse, Thematizer is fed one or multiple texts that the user has entered. Each text is first cleaned of textual and clerical noise (i.e., white spaces and punctuation) before any thematic parsing occurs. Next, the overall themes and rhemes for each sentence are identified and extracted. Any marked themes identified parallel to overall themes are then classified into their marked theme type and semantic class. The thematic progression patterns for each sentence are determined afterwards. Finally, the analytical results from these parsing tasks are output to the web interface for data visualization.

As the linguistic analyses programmed for each parsing task embody the core functionality of Thematizer, each step is explained in detail in its respective sections below.

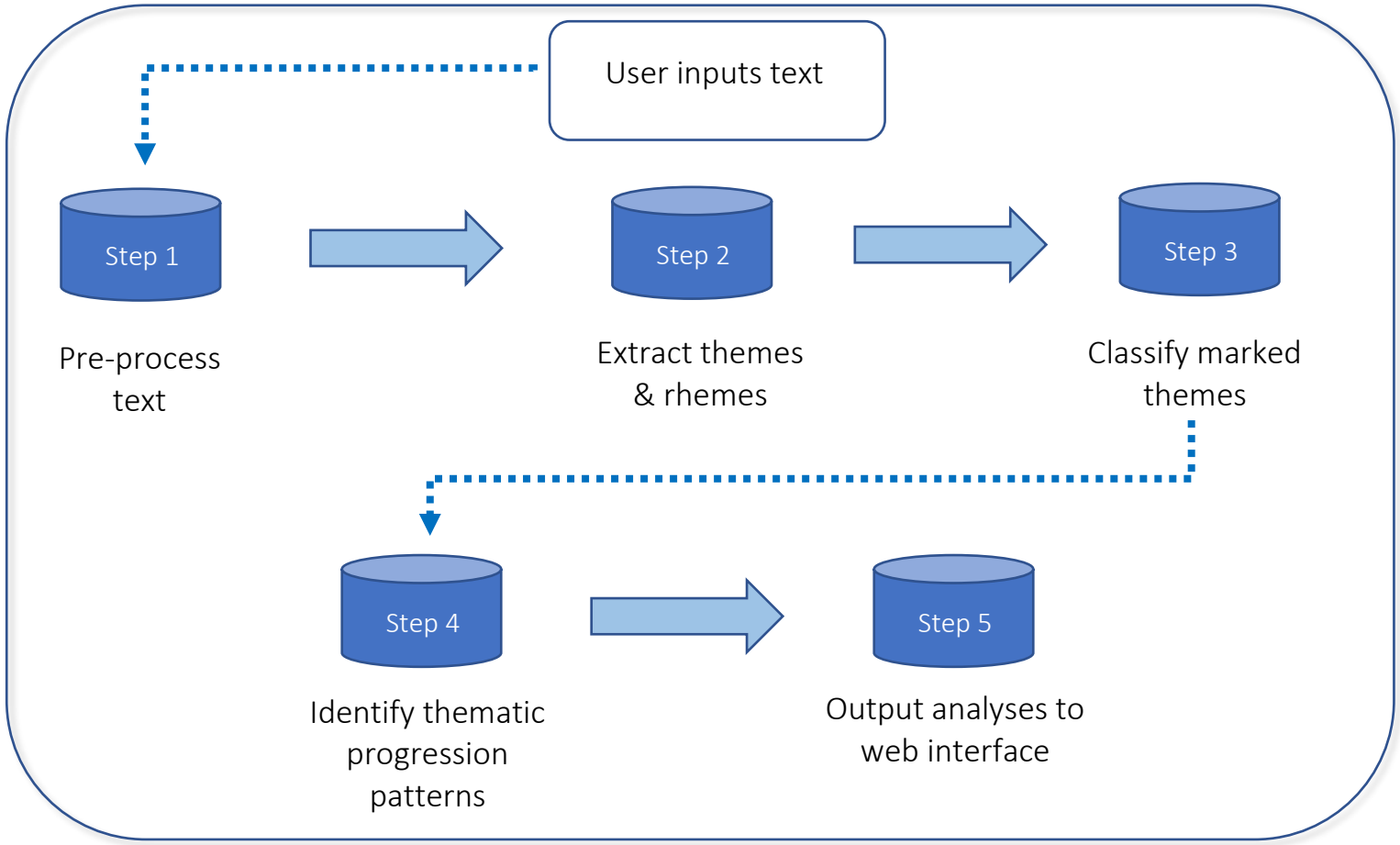


Figure 4-17: Core processing steps that Thematizer progresses through for text analysis. The three core thematic parsing steps correspond to steps two to four, where thematic constituents are identified, extracted, classified and analyzed in terms of thematic progression.

4.7.1 Text Pre-Processing

Before Thematizer progresses through the three core thematic parsing steps, the input text undergoes cleaning via pre-processing. Pre-processing the text is a critical step in facilitating an accurate analysis of the input with as little cotextual noise as possible. Such noise is nearly entirely dependent on the text type, which is why accounting for typical punctuation, structural and formatting characteristics of the text type during pre-processing is indispensable. In fact, the overall accuracy of the intermediary dependency and final thematic parses directly rely upon how clean the input is.

The first pre-processing step regardless of text type is identifying newlines or line breaks from the text input. These are of particular importance as they not only signal a new sentence to the parser but also section headers, titles and headings. Wherever a newline or line break occurs, Thematizer automatically adds a period to force a sentence boundary. This prevents Spacy from including section headers and titles as part of the subsequent sentence. Inserting periods is also important for lyrics, which typically lack punctuation marks to indicate sentence boundaries. Newlines also come into play when bullet points are present. If no punctuation is present at the end of a clause within a bullet point or numbered list, then a period is added to separate it from subsequent bullet points.

Once the text has been cleaned of preliminary clerical noise, sentences are split into their so-called **t-units**. In thematic analyses, t-units are the basic unit of text. These include the independent matrix α-clause and any dependent hypotactic β-clauses; t-units are thus "beta [dependent] clauses, including their thematic structure, [which] tend to be constrained by the alpha [independent] clauses" (Leong 2019: 3). Specifically, t-units contain the matrix α-clause's root verb, its congruent subject and any hypotactically instantiated clauses. Example (1) in Figure 4-18 illustrates hypotaxis (dependent β-clause + independent α-clause) as a single t-unit together with the matrix clause.

SINGLE T-UNIT	
β-clause	**α-clause**
(1) *While statistics are not available*	*identical bio-performance for a BCS II compound was likely to be low.*

Figure 4-18: Sentences comprised of a fronted hypotactic β-clause and followed by the matrix α-clause are parsed as a single t-unit.

Here, the hypotactic clause is introduced with the adverbial *while* to form the β-clause. This is combined with the subsequent independent α-clause to constitute the entire t-unit. The theme of the single t-unit would therefore be: *While statistics are not available, identical bio-performance for a BCS II compound*.

T-UNIT 1		T-UNIT 2
α-clause 1	**β-clause**	**α-clause 2**
(2) *Shen et al. have found that*	*the resistivity of superconducting materials presents a nonlinear behavior*	*but no general model to describe or predict the resistance variation has been published to date.*

Figure 4-19: Coordinated sentences with multiple matrix α-clauses and corresponding subjects are split into two separate t-units and parsed individually.

Conversely, paratactic sentences are traditionally split into two t-units since they, by nature, are two matrix α-clauses connected via coordinating conjunctions or conjunctive adverbials. The coordinating conjunction *but* in (2) from Figure 4-19 splits the first α-clause starting with *Shen et al.* from the second α-clause starting with *no general model*. Thus, despite syntactically

forming a single sentence, for the purpose of thematic progression analysis, (2) contains two t-units. Thematizer splits these two clauses into two separate sentences and inserts a period between the two. This allows both to be analyzed in terms of their respective themes, rhemes and thematic progression.

SINGLE T-UNIT	
α-clause 1	**α-clause 2 with elided subject**
(3) *The external quality factor depends not only on the resonator characteristics, but also the coupling elements,*	*and can be calculated in case of a capacitive, inductive, and both capacitive and inductive couplings.*

Figure 4-20: Coordinated sentences with multiple matrix α-clauses are parsed as one t-unit if the same subject of the first α-clause is elided in the second α-clause.

Coordinated sentences whose subject has been elided in the second paratactic clause form a single t-unit, as exemplified in Figure 4-20. There, the grammatical subject *external quality factor* is elided in the second α-clause, which would imply constant thematic progression through instantiation of the same theme. Justification for considering such instances as a single t-unit becomes clear when erroneously and unnecessarily separating the two α-clauses from one another. In doing so, the second α-clause would become its own t-unit *And can be calculated in case of a capacitive, inductive, and both capacitive and inductive couplings*, which lacks a grammatical subject and thereby theme. The new t-unit would therefore become entirely rhematic, which was already the case before the original sentence was split. Thus, parsing a paratactic sentence whose second subject was elided as a single t-unit with two α-clauses not only maintains thematicity, it also obviates an additional parsing step during text processing.

SINGLE T-UNIT	
β-clause	**α-clause**
(4) *If the criminality of a particular act can only be adjudged after the fact and there are possible grounds for excluding criminal responsibility,*	*then how can we deter perpetration of the act in the present?*

Figure 4-21: If parataxis occurs within the β-clause, it is not split into a separate t-unit but remains a compound clause embedded within the β-clause. This is then parsed as a singular t-unit together with the α-clause it is dependent upon.

In Figure 4-21, a paratactic clause is inserted within the β *if*-clause. Since the coordinated *there are possible grounds for excluding criminal responsibility* remains dependent due to its realization within the hypotactic adverbial *if*, a complex hypotactic structure emerges. Regardless, (4) is still considered a single t-unit on account of its β-clause + α-clause definition. Further justification for this approach is the inability to shift the paratactic phrase *and there are possible grounds for excluding criminal responsibility* to the end of the sentence. Doing so would result in a semantically incoherent statement as well as incongruity between the question formed in the main clause and the newly shifted declarative clause.

Additional cases of t-units to be split are consecutive independent clauses separated by a punctuation mark (specifically, a colon, semicolon or hyphen) and/or a conjunctive adverbial. Examples of the latter are *however, thus, conversely*, *thereby,* and *consequently*.

An example of two α-clauses conjoined with a semicolon and a conjunctive adverbial can be seen in Figure 4-22,

T-UNIT 1	T-UNIT 2
α-clause 1	**α-clause 2**
(5) *The experience was of little consequence;*	*however, certain aspects remained important.*

Figure 4-22: Sentences that contain two α-clauses conjoined by a conjunctive adverbial, semicolon or hyphen are split into two separate t-units and parsed individually.

where *however* functions as the conjunctive adverbial. Similar to (2) in Figure 4-19 above, the two independent clauses are split into two t-units and the punctuation mark is replaced with a period. The resulting t-units would then be *The experience was of little consequence* and *However, certain aspects remained important.* The period is inserted so that Spacy can correctly identify the end of the sentence boundary; otherwise, Spacy would have considered two independent clauses joined by a punctuation mark or conjunctive adverbial as a single sentence. Splitting the original single sentence into two t-units allows tracing thematic and rhematic elements across both t-units, which would have otherwise remained rhematic alone in the original.

SINGLE T-UNIT	
α-clause 1	**α-clause 2**
(6) *The experience was of little consequence,*	*certain aspects remained important.*

Figure 4-23: Sentences that contain two α-clauses conjoined by a comma alone are treated as a single t-unit.

If two independent sentences are strung together with commas alone but *without* conjunctive adverbials or coordination, they are considered run-ons and are not split in pre-processing (cf. Figure 4-23). The reason for this is the amount of dependency parsing required to determine sentence boundaries. Since run-ons can consist of concatenated independent and subordinate clauses, the requisite dependency testing would add considerable processing time. As run-ons themselves can throw the thematic progression of a passage into disarray, attempts to dissect them would only compound subsequent parsing issues. Despite potential misparses in run-ons, the highlights in the analytical output from Thematizer could serve as a visual indication to the user that run-ons are present since the identified themes would erroneously span beyond their standard boundaries. For example, in the output for (6) in Figure 4-23, Thematizer would identify the theme as *The experience was of little consequence, certain aspects*. Since that thematic span exceeds the first grammatical subject and congruent finite verb *was*, such output could be indicative of potentially poor grammatical structure. This could then prompt the user to maintain the run-on for stylistic purposes or remove it for conventional grammaticality purposes.

T-UNIT 1	T-UNIT 2
β-clause	**α-clause**
(7) *"There are more apps than ever before that can give you more control over your finances,"*	*Grant says.*

Figure 4-24: Sentences that contain a direct quotation as a β-clause and a projecting matrix α-clause are split into two separate t-units and parsed individually. The direct quotation must fulfill syntactic independence when extracted from the sentence to merit the split.

The final case of potential t-unit identification is in sentences containing quotation marks, as shown in Figure 4-24. These form a unique case in that they can introduce either dependent or independent clauses in the form of direct quotes. It is the latter that proved most important in

the development of Thematizer since, if not separated as an independent and new sentence, the software would consistently misparse the original.

Spacy generally marks the finite verb in the independent matrix clause as the verbal root of the sentence, the token *says* from the α-clause in Figure 4-24. This causes Thematizer to parse everything up to the verbal root as theme when two separate themes and rhemes are in fact present. To circumvent this parsing issue, the pre-processor divides the original sentence into two: one before and one after the direct quote. Note that if quotation marks are not present and an indirect quote is referenced with projecting clauses such as *he said/stated/argued/claimed that*, then the pre-parser ignores this; such instances are then parsed as projecting themes so long as the *that*-conjunction is present. If the clause within the quotation marks is dependent, then it is not split from the matrix clause and the sentence remains one t-unit, as is the case in Figure 4-25.

SINGLE T-UNIT
α-clause

(8) *Other names used are 'tone unit' or 'intonation unit.'*

Figure 4-25: Sentences containing quoted elements remain a single t-unit if such elements are syntactically dependent on the matrix α-clause and do not fulfill syntactic independence when extracted from the sentence.

The last pre-processing step involves removing extraneous, exceptional or any remaining punctuation, some of which may have resulted from splitting t-units. Often, text may contain clerical errors such as recurring punctuation marks in sequence or extraneous whitespaces (e.g., repeated commas as in *He bought milk,, but no bread*). While the latter rather affects the layout and presentation of the text alone, the former can have greater consequences for the parsing itself. Parsers that have been trained to detect sentence boundaries, as is the case with Spacy, rely upon cotextual cues to accurately split texts. Extraneous or erroneous punctuation can readily impair that functionality. Therefore, the pre-processor addresses such cases by removing problematic punctuation. Further, removing any forms of parentheses, brackets, slashes and any text in between these punctuation marks reduces the likelihood of errors from the parse. This does come at a cost to information retention, however, whereby parenthetical, anecdotal or additional information provided between the removed parenthesis types is elided as well. The frequent use of parentheses and brackets is particularly common in scientific or academic texts, wherein long lists of references are cited. Despite this potential loss of secondary information, the cleaned text is less likely to throw parsing errors due to punctuation-induced noise.

Once the text is cleaned, it is ready for subsequent part-of-speech tagging and dependency parsing, both of which form the foundation of the remaining linguistic analyses.

4.8 Identifying Theme and Rheme Spans in Text

The first core thematic parsing step that Thematizer undertakes is the identification of the overall theme and rheme spans of each sentence in the text (as exemplified below in Table 4-4). These include marked themes, the grammatical theme and the rheme. These spans are critical for parsing marked themes and each sentence's thematic progression pattern in subsequent steps. This section therefore outlines how Thematizer makes use of Spacy's dependency parses and the tokens' indices for theme and rheme extraction. For clarity's sake, the index is the numerical reference point that Spacy uses to identify the location of a token within a text.

First and foremost, the parser attempts to identify the so-called root of the sentence. For Spacy, each sentence/clause may have one and only one root. In independent clauses, the finite verb is

always marked as the sentence's root.[20] If a verbal root is found, then Thematizer stores its index as a potential boundary marker for the beginning of the rheme (cf. index three in Table 4-4). With the index of the verbal root, the index of the congruent subject is identified, which equates to index one in Table 4-4. This assumes that there is a congruent subject and that Thematizer has identified one as the dependent of the root. If no subject or verbal root is found, then the entire sentence is marked as a rheme.

Assuming the indices for the finite verb and congruent subject have been identified, Thematizer marks all text up to the matrix subject's index and dependents as the theme and all constituents thereafter as the rheme. Note that the verbal root of the matrix clause is required to determine the corresponding grammatical subject, which is dependent on the root. This ensures that subject-verb pairs in subordinate clauses are not identified as the matrix' subject-verb pair.

	THEME		**RHEME**					
Text	*Morality*	*stories*	*typically*	*espouse*	*common*	*Christian*	*values*	.
Dependency	COMPOUND	NSUBJ	ADVMOD	ROOT	AMOD	AMOD	DOBJ	PUNCT
Index	0	1	2	3	4	5	6	7

Table 4-4: Identification of theme and rheme boundary via dependency and index parsing without right dependents.

There are cases, however, where the matrix subject's index is insufficient in delineating the theme's boundary to the rheme. This is illustrated in Table 4-5, whereby the right dependent *from Chaucer* follows the subject. Although the prepositional phrase appears after *morality stories*, it is dependent on the subject and therefore must be parsed as belonging to the theme. The resulting theme then spans from index 0 to 3, with the rheme spanning indices 4 to 8. In this work, matrix subjects not only form the overall theme but are also denoted as the grammatical theme as inspired by Weis' terminology of the psychological, i.e., grammatical, subject. Grammatical themes are always considered unmarked and stand in contrast to marked themes, which are sentence constituents that appear before the grammatical subject (cf. Chapter 4.3). How marked themes are parsed will be addressed in the next section, but it should be noted that their parse occurs parallel to grammatical theme and rheme identification outlined here.

	THEME				**RHEME**				
Text	*Morality*	*stories*	*from*	*Chaucer*	*typically*	*espouse*	*Christian*	*values*	.
Dependency	COMPOUND	NSUBJ	PREP	POBJ	ADVMOD	ROOT	AMOD	DOBJ	PUNCT
Index	0	1	2	3	4	5	6	7	8

Table 4-5: Identification of theme and rheme boundary via dependency and index parsing with right dependents.

Both examples in Table 4-4 and Table 4-5 represent standard, default word order without any marked themes. In the formal parsing order defined in Thematizer's code, these cases are tested last but were explained here first to provide an understanding of the relationship between dependency, index and theme/rheme span.

Thematizer first tests for sentence structures that require exceptional treatment, namely interrogatives, existentials and clefts. When processing interrogatives, Thematizer must first determine whether the text in question is declarative or interrogative in mood (cf. Table 4-6). The latter is ascertained through the presence of a WH-interrogative, inversion of the subject and finite verb and a question mark as a sentence-final punctuation mark. The question mark and subject-verb inversion ensure that embedded questions are not parsed as an interrogative but rather declarative sentences. If a WH-interrogative is found within the interrogative being

[20] Note that Spacy identifies other sentence constituents as the root if no verbal root is found, such as nominal or adjectival constituents, as may be the case in phrases or clauses without a verb.

parsed, the entire simple or compound interrogative is extracted. With polar interrogatives, the root's congruent subject and any right dependents are identified with their indices to determine the corresponding theme span. The remaining indices and tokens after the theme span are then relegated to the rheme.

	THEME			RHEME			
Simple WH-Interrogative	*Why*			*did*	*they*	*answer*	*?*
Index	0			1	2	3	4
Compound WH-Interrogative	*How*	*long*	*ago*	*did*	*it*	*happen*	*?*
Index	0	1	2	3	4	5	6
Polar Interrogative	*Can*	*you*		*really*	*believe*	*that*	*?*
Index	*0*	*1*		*2*	*3*	*4*	*5*

Table 4-6: Theme and rheme spans in simple and compound WH-interrogatives as well as polar interrogatives with their corresponding indexical spans.

Clefts form the next exception cases that Thematizer may be confronted with. These are exceptional both in how they are parsed thematically and in the function they fulfill rhetorically (cf. Chapter 4.3). Beginning with non-projecting clefts, these must contain a dummy-*it* and are followed by a copula, adjective and infinitive phrase in its basic form, as shown in Table 4-7. It is worth reminding that clefts function to emphasize the propositional content after the *it is* cleft structure. The *it is* is semantically void but pragmatically salient as a signal to the importance of the information to come. It therefore constructs an established rhetorical basis, i.e., as an emphatic or attention-drawing marker, upon which the following rheme expounds. On account of this rhetorical function and lacking coreferentiality with the dummy-*it*, the entire *it is* phrase is considered the grammatical theme.

	THEME		RHEME					
Cleft	*It*	*is*	*critical*	*to*	*measure*	*their*	*full*	*experience*
Dependency	NSUBJ	ROOT	ACOMP	INF	XCOMP	POSS	AMOD	DOBJ
Index	0	1	2	3	4	5	6	7

Table 4-7: Theme and rheme spans in non-projecting clefts with corresponding dependencies and indices.

Since Spacy is unable to distinguish between a coreferential and non-coreferential *it*, Thematizer employs Spacy's pattern matching functionality to isolate cleft structures. This structure was generalized such that Spacy searches for *it* as the grammatical subject of a matrix followed by any form of the copula *to be*. It was important to allow for any form of *to be*, such as *may have been*, since auxiliaries and modals may take the place of the generic *is* in the cleft structure. Additionally, dependency parses ensured that *it* was both the grammatical subject of the matrix clause and thereby the dependent of the matrix' finite root. Without this condition, any instance of *it is* within a span of text would have been identified. Finally, an adjective with the universal dependency of adjectival complement (ACOMP) or attribute (ATTR) and subsequent infinitive phrase completed the generalized search pattern.

When the input text is passed on to Spacy for initialization, any stretches of text whose dependencies and textual realization match the cleft pattern are then marked. Once Thematizer reaches a sentenced marked for a cleft, it circumvents default theme parsing by extracting the indices corresponding to *it is* or variations thereof. The text span of these indices is then saved as the grammatical theme with the remainder defined as the rheme.

Along the same lines as clefts, existentials, too, function as a rhetorical marker. Just as the dummy-*it* is non-coreferential, the existential *there* is non-deictic and thereby non-coreferential. Since Spacy returns the dependency marker EXPL for instances of existentials, Thematizer can use the corresponding index of this dependency together with the verbal root index to extract the theme span (cf. Table 4-8). Again, it was important that the existential was linked to the verbal root of the matrix clause to prevent existentials within subordinate clauses from being erroneously extracted. The only exception to this was in projecting themes, as will be outlined in the next sections on marked theme extraction and classification.

	THEME			RHEME				
Existential	*There*	*have*	*been*	*similar*	*advances*	*in*	*computer*	*tasks.*
Dependency	EXPL	AUX	ROOT	AMOD	ATTR	PREP	COMPOUND	POBJ
Index	0	1	2	3	4	5	6	7

Table 4-8: Theme and rheme spans in existential structures with corresponding dependencies and indices.

In summary, for each sentence parse, Thematizer determines the verbal root and grammatical subject of the matrix clause, queries for interrogative or declarative form, and identifies any potential cleft or existential structure. If no verbal root or grammatical subject was found, then the entirety of the sentence becomes the rheme as is commonly the case for section headers or imperatives, for instance. Should both subject and finite verb dependencies be found, then their indices are stored as boundary markers for the theme and rheme spans of the sentence. Once the indexical spans have been identified, these are saved as so-called entities, using Spacy terminology, for quick access to the text spans in later analyses. Instead of having to store the exact indexical spans, the entities THEME and RHEME can be extracted for each sentence, whose textual realization and index can then be accessed directly from each entity.

While it is not uncommon for sentences such as those outlined above to appear, most reveal greater complexity through compound subordination and a mixture of clause types within a single sentence. Therefore, the methodology for grammatical themes and rhemes described above should not be considered a static case; interrogatives mixed with existentials or clefts, polar interrogatives with right dependents and grammatical themes and both left (prepositioned) and right (post-positioned) dependents are also captured by the parsing approach.

Thematizer's first core thematic parsing step involves the identification of the finite verbal root of the matrix clause, which functions as the boundary between theme and rheme spans. Dependency parses and indices are used in order to delineate the individual spans of all thematic constituents to then be saved as entities for later retrieval and analyses. If marked themes are present in front of the grammatical theme, their textual span is also separated from grammatical themes. This marked theme span is then passed on for subsequent marked theme classification and semantic categorization, which is Thematizer's second core thematic parsing step.

4.9 Marked Theme Parsing

Marked theme parsing constitutes the second pillar of Thematizer's thematic analyses and is split into two tasks: marked theme extraction and marked theme classification. In the first step, all marked themes in the sentence, of which there may be more than one, are separated from each other and their entire textual span is extracted via dependency and index parsing. In the second step, the extracted marked themes are classified into their marked theme type (e.g., circumstantial or structural) and then into their semantic subclass (e.g., TEMPORALITY or CONCESSIVE). The parsing requirements and steps for marked theme extraction are outlined in Figure 4-26 and Chapter 4.9.1 with marked theme classification in Chapter 4.9.2.

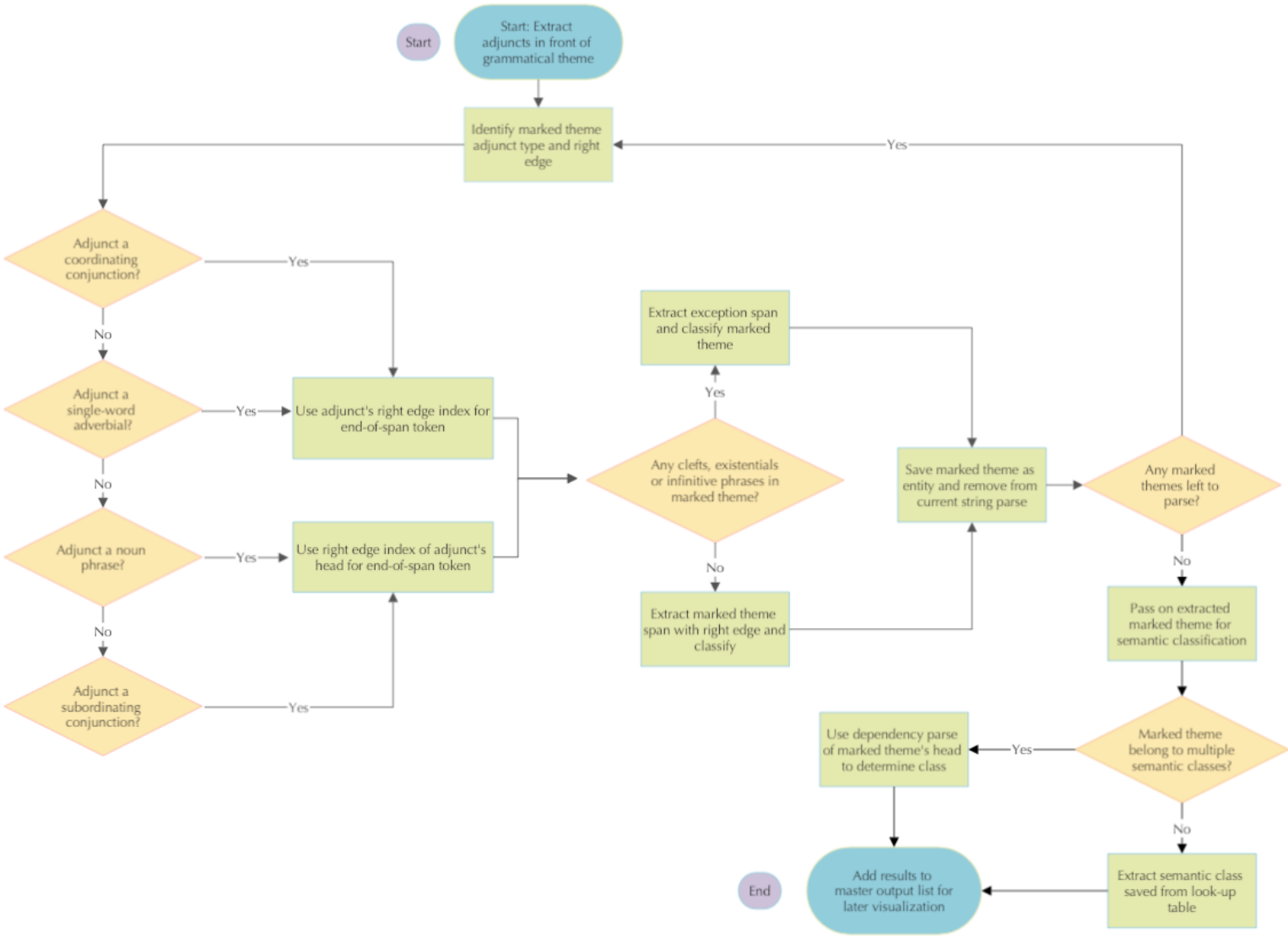


Figure 4-26: Breakdown of individual processing steps that Thematizer performs when extracting and classifying marked themes.

4.9.1 Marked Theme Extraction

Adjuncts, by definition, are optional sentence constituents that can take the form of circumstantial adverbials (*in 1982*), structural adverbials (*as such*), modal adverbials (*amazingly*), and hypotactic adverbials (*After testing*). When adjuncts appear sentence initially, Thematizer categorizes them tentatively as a marked theme for subsequent parsing. This parsing takes place parallel to the identification of the grammatical theme and rheme as outlined in Chapter 4.8.

Once found, the entire marked theme passage is first extracted from the sentence being parsed, which may contain more than one marked theme, as shown in Table 4-9. Therefore, when parsing marked themes, Thematizer must determine how many marked themes are present and what the indexical spans of the marked themes are. In Table 4-9, the marked modal theme *unsprisingly* spans index 0 alone; conversely, the hypotactic marked theme *when the reviews are positive,* spans indices 1 to 6. How Thematizer identifies the number and span of marked themes is illustrated and will be explained in detail in the following.

	MARKED THEME							THEME		RHEME		
Thematicity	MODAL THEME	HYPOTACTIC THEME						GRAMMATICAL THEME		RHEME		
Text	*Unsur-prisingly*	*when*	*th e*	*revie ws*	*ar e*	*positive*	*,*	*the*	*agent*	*feel s*	*validate d*	*.*
Index	0	1	2	3	4	5	6	7	8	9	10	11

Table 4-9: Identification and extraction of multiple marked themes within the theme span on the initial basis of the marked themes' indices, which are later used with dependency parses to extract the exact marked themes from the sentence

When Thematizer receives a text passage denoted as a marked theme, it first assumes that the sentence-initial token initializes the marked theme span. While marked themes can be single-word adverbials such as *unsurprisingly* in Table 4-9, they can also be comprised of multiple sentence constituents, such as the hypotactic theme from the same table. Therefore, Thematizer must determine where the marked theme begins and ends. In cases where multiple marked themes are present in a text to be parsed, simply using the final index of the text passage would result in erroneous marked theme extractions. In Table 4-9, for example, identifying the marked theme span from indices 0 to 6 would cause Thematizer to return *unsurprisingly when the reviews are positive* as the entire modal theme and thereby ignore the hypotactic marked theme. Instead, *unsurprisingly* should be extracted as a single marked modal theme with *when the reviews are positive* as the second marked hypotactic theme.

How Thematizer initially achieves this is with the so-called **right edge index**, also called the **right dependent**. This index refers to the furthermost right dependent of a token and can be used to identify the ending index of a marked theme span. A schematic breakdown of this indexical and dependency analysis is shown in Figure 4-27 below. There, the right edge of *unsurprisingly* refers to index zero, which is the adverbial *unsurprisingly* itself. In the case of single-word marked themes, then, the right edge index is the index of the adverbial itself. Where marked themes span entire clauses, as is the case with *when the reviews are positive*, the right edge index of the adverbial is also the adverbial itself, *when*. However, this is not the end of the marked theme span. In such cases, **the right edge of the marked theme's head** is required. In this example, the syntactic head of *when* is the finite verb *are*, whose right edge index is then *positive*, i.e., the end of the marked theme span. This differentiation is of particular importance in hypotactic structures, which have their own grammatical subject and finite verb. Similarly,

in compound adverbials, e.g., *more precisely*, the head of the clause-initial token *more* is the second adverbial in the compound that concludes the compound span, i.e., *precisely*. Using the right edge of the marked theme's head thereby ensures that the entirety of the marked theme span is accounted for.

Once Thematizer identifies and extracts a marked theme span, it is concatenated to an intermediary list. The extracted marked theme is removed from the text passage string and an additional parsing condition tests whether the string is then empty. An empty string indicates that all marked themes have been processed. Otherwise, the remaining text to be parsed in terms of residual marked themes is passed on for additional analysis. In the example from Figure 4-27, the modal theme *unsurprisingly* is identified and extracted using its right edge index. This marked theme is added to the intermediary list of marked themes, which leaves the second marked theme *when the reviews are positive* to be parsed. Thematizer then receives this text passage alone and follows the same steps outlined above to extract the marked theme span in its entirety by using the right edge index of the head of *when*.

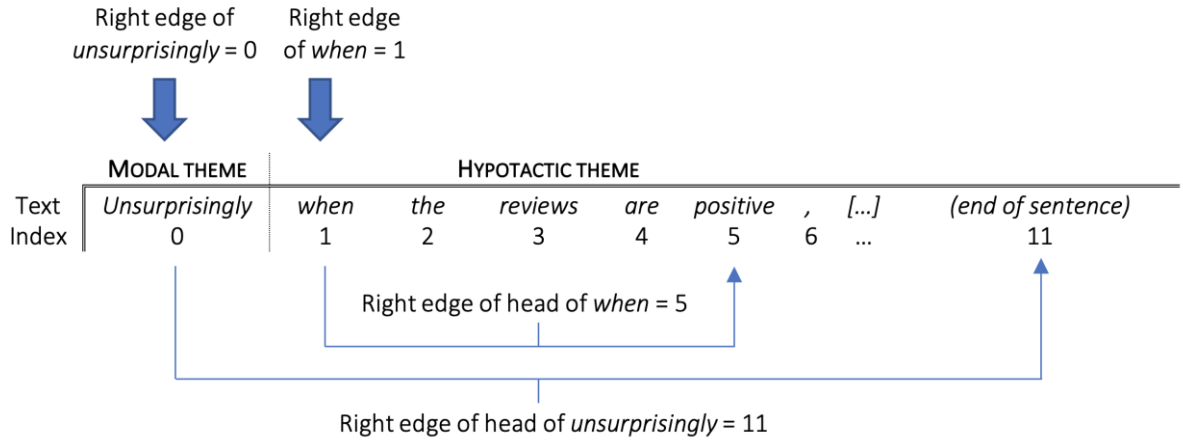


Figure 4-27: Visualization of right edges and the right edge of the marked theme's syntactic head as indexical indicators of where the marked theme span ends and as used for marked theme extraction.

Thematizer thus accounts for any number of marked themes in a text passage through recursive parsing, i.e., marked theme parsing until all marked themes have been extracted. This recursive step is evident in the flowchart from above in the decision diamond querying 'Any further marked themes to parse?' and exemplified textually for further clarification in Table 4-10. In the first step, the entire text string from the initial theme/rheme parse is passed on for marked theme parsing. Thematizer extracts the first marked theme *However* with its right edge, removes the span from the original text string, and checks whether any text remains. Since the string is not empty but rather still contains the remaining four marked themes to parse, the marked theme parsing function is called again with the new text string. This continues through all five marked themes, as indicated by the five parsing passes, and concludes once the final marked theme *amazingly* is removed from the string. This causes the string to become empty, which terminates marked theme parsing.

	Text String to Parse	Marked Theme Extracted	String to Pass On
Parsing Pass 1	*However, in the end, when examined closely, results indicate that amazingly*	*However*: Structural theme	*in the end, when examined closely, results indicate that amazingly*
Parsing Pass 2	*in the end, when examined closely, results indicate that amazingly*	*in the end*: Circumstantial theme	*when examined closely, results indicate that amazingly*
Parsing Pass 3	*when examined closely, results indicate that amazingly*	*when examined closely*: Hypotactic theme	*results indicate that amazingly*
Parsing Pass 4	*results indicate that amazingly*	*results indicate that*: Projecting theme	*amazingly*
Parsing Pass 5	*amazingly*	*amazingly*: Modal theme	EMPTY STRING: END PROCESSING

Table 4-10: Illustration of recursive parsing steps Thematizer proceeds through when it receives a text string with multiple marked themes to identify indexically and extract individually.

Aside from the conventional cases of marked theme processing presented thus far, there are two exception cases that Thematizer may be confronted with at this point in the parse. These exceptions include projections and post-positioned adverbials. Neither of these cases can employ the approach outlined above on account of their syntactic structure. Projections form their own class of marked themes and are typified through their use of the subordinating conjunction *that* or *whether* following a projecting matrix clause. The four types of projections are illustrated in Table 4-11.

When text with projections is passed on for marked theme parsing, Thematizer searches for the dependency MARK, the part of speech SCONJ and the corresponding textual realization *that* or *whether*. If all three conditions are found within the text passage, then the parser considers it a projecting theme. If the projection contains an adjective complement (as with *clear* in (1)), then it is classified as an adjectival projection. Should the subject of the projecting clause not be a dummy-*it* and the finite verb fall into a set of pre-defined interpersonal verbs, then Thematizer classifies the projection type as interpersonal projection. The collection of pre-defined verbs was motivated by the work done by Rudolph & Försterling (1997).

Thematicity	PROJECTING THEME	GRAMMATICAL THEME	RHEME
Adjectival Projection	*(1) It is clear that*	*the work*	*remained unfinished.*
Interpersonal Projection	*(2) Brooks believes that*	*the work*	*remained unfinished.*
Objectifying Projection	*(3) These are indications that*	*the work*	*remained unfinished.*
Experiential Projection	*(4) The results found that*	*the work*	*remained unfinished.*

Table 4-11: Thematic analysis of the marked theme type projecting theme *and the four realizational types that projections can possess.*

Next, to be qualified as an objectifying projection, the subject of the projecting matrix clause may not be a dummy-*it* and the predicate must be comprised of a copula and a noun phrase. While this structure resembles that of a relative clause, e.g., *That is a book that I read*, the critical difference is the *that* dependency: should *that* function as a relative clause, then it is does not have a MARK dependency, which would be the case for a projecting *that*. Therefore,

this dependency requirement automatically excludes cases of *that* functioning as a relative pronoun. Finally, if none of the previous conditions for projection classification are fulfilled, then Thematizer defaults to an experiential projection.

The second exception case that Thematizer may need to address is that of post-modifying adverbials. These form a closed class of adverbials that appear after the word they are describing, and are limited to instances such as *notwithstanding, prior, later, ago, before* and *away.*[21] Examples of each adverbial used in context are found in (5) – (7). Due to the closed-class nature of these adverbials, their instantiation in the marked theme text is searched for by using string-based pattern matching. That means that the adverbial itself is used as a search parameter within the marked theme passage instead of the right edge index. If *notwithstanding* is found within the passage, then post-modifying adverbial parsing ensues. Since these adverbials mark the end of the clause already, the marked theme span can be extracted with the clause-initial index and the post-modifying adverbial's index. This approached obviates the need for any dependency or right edge parsing while ensuring correct extraction of the span.

	MARKED THEME	THEME	RHEME
Thematicity	CIRCUMSTANTIAL THEME	GRAMMATICAL THEME	RHEME
(5)	*The weather notwithstanding,*	*the trip*	*was a success!*
(6)	*Three days prior/later/ago/before,*	*there was*	*a mess.*
(7)	*Far away,*	*we*	*saw some mountains.*

A final case to mention is the insertion of adverbials within marked themes, whereby one marked theme is inserted within another. In (8), the hypotactic marked theme in bold contains the underlined structural theme *in fact*. Due to how marked themes are parsed with right edges, marked themes inserted within others as found in (8) cannot be accounted for. If concatenated linearly, i.e., interstitially, as was the case in Table 4-10 above, then Thematizer can treat each marked theme individually.

(8) **When the tests, <u>in fact</u>, reached peak efficiency**, the required processing power at that point was measured.

The programmatic reason for this non-insertional approach is due to how Spacy defines entity spans. Parsing inserted marked themes would additionally result in the splitting and thereby doubling of marked theme entities, as shown in Table 4-12. Spacy does not allow entities to overlap since indexing errors would occur when referencing the corresponding entity's indexical span.

	HYPOTACTIC THEME A				STRUCTURAL THEME			*HYPOTACTIC THEME B
Text	*when*	*the*	*tests*	*,*	*in*	*fact*	*,*	*finished*
Index	0	1	2	3	4	5	6	7

Table 4-12: Erroneous splitting of the superordinate hypotactic theme into two separate hypotactic themes on account of the inserted structural theme. The asterisk before hypotactic theme B indicates the erroneous split that would result in misidentified indices in marked theme entities.

For that reason, the original hypotactic theme *when the tests, in fact, finished* would be overwritten with the inserted structural theme *in fact*. This would result in two separate

[21] This list is by no means exhaustive but was collated during the testing process. In future developments of the program, additional post-modifying adverbials will be added to the list once found.

hypotactic themes, one spanning from index zero to three, the other spanning index seven only. Therefore, where only one hypotactic theme was actually present, Thematizer would have incorrectly output two. This potential parsing result further motivated the decision to extract superordinate marked themes alone without separating embedded marked themes.

Inserted marked themes could have been accounted for by using an alternative parsing scheme and syntactic requirements. In fact, in the first version of Thematizer, adverbial phrases were extracted on a lexical basis, not a syntactic, phraseological one. While that allowed for identifying interstitial adverbials, it came at the expense of efficiency: many adverbial phrase spans were misparsed due to incongruent index matches, which caused Thematizer to crash entirely. Otherwise, lexically parsed adverbial phrases occasionally caused endless recursion to occur. This meant that the parser incorrectly identified an adverbial phrase, e.g., first extracting *in* from *in the car* and assuming that *the car* formed the next marked theme. Since no matching adverbial was found within the phrase *the car*, the marked theme parsing was executed *ad infinitum*, which ultimately caused Thematizer to crash. As the phraseological and right edge approach eliminated cases of endless recursion and erroneous breakdowns of already parsed marked themes, a lexically based parsing approach was deemed inappropriate and insufficient.

Once the marked theme(s) have been extracted, they are ready for classification into their marked theme and semantic class type, which is the second and final parsing step for marked themes before being stored for data retrieval and output.

4.9.2 Marked Theme Classification

Once a marked theme span has been extracted and separated from any other marked themes present, it is passed on for classification to determine its marked theme type and its semantic subclass. Marked themes fall into five separate categories, namely modal theme, circumstantial theme, structural theme, projecting theme and hypotactic theme (cf. Chapter 4.3 for an explanation of which syntactic and textual patterns inform their marked theme type). These marked theme types can then be broken down further into their semantic subclass, some of which are TEMPORALITY, CAUSALITY, ADDITION and ANGLE. The purpose of this section is to outline how Thematizer accomplishes both classifications.

As elucidated in the previous section, marked themes can be realized in a myriad of ways and at varying token lengths. Structural themes, for example, are a collection of adverbial phrases that establish cohesion across sentences, e.g., *therefore* or *yet*. Hypotactic themes, too, are a collection of adverbial phrases that begin with a subordinating conjunction and introduce a dependent clause through tokens such as *because* or *after*. Circumstantial themes are the most flexible with greater room for open-ended classification, however, their realization patterns can also be drawn back to either prepositional, adverbial or temporal noun phrases. Examples are *in the book*, *quickly* or *three days prior*. Here, one marked theme consists of a single word alone while others span entire phrases. Thematizer thus must be able to identify the characteristic word or phrase that instantiates the marked theme for corresponding classification.

To do so, a list of adverbials, prepositions, subordinating and coordinating conjunctions and temporal noun phrases was collated from the 150 training texts, English grammar books and previous research on adjuncts (e.g., Huddleston et al. 2021: 208-228; Ma & Zhu 2023: 467-468; and Halliday & Matthiessen 2014: 108-109). Once collected, these were categorized according to their marked theme type. For instance, the modal adjunct *in fact* was added to the collection of modal themes; the phrase *in the case of* to the collection of circumstantial themes; and the adverbial *when* to the collection of hypotactic themes. Each word or phrase was saved

together with its length in tokens and its semantic subclass (e.g., *when* of token length one and belonging to the semantic class of TEMPORALITY). Altogether, this information was saved as an individual CSV file for each marked theme class, which functioned as look-up tables during marked theme parsing. With these look-up tables, the first step to marked theme classification could begin.

Once Thematizer receives a marked theme to be classified, it first determines the text's length in tokens. It was recognized early in development that the length of a marked theme could be used as a parameter to speed up processing. Instead of searching through the entire look-up table for each marked theme, the token's length could set a cut-off point for marked themes that Thematizer should search for. For example, the text in (1) has the two marked themes *contrarily* and *when in doubt* marked in bold. Since the first marked theme only has a length of one, Thematizer ignores all marked themes saved in the look-up tables with a token length greater than one. That then reduces the number of patterns to search through from 403 total marked themes to 212.

(1) **Contrarily**, **when in doubt**, you should consult the ledger!

For *when in doubt*, with a token length of three, 353 patterns are searched through instead of all 403. Thematizer first searches for a matching marked theme with three tokens. If no match is found, Thematizer moves on to marked themes with two tokens. Should no matches be returned with two tokens, then it searches for marked themes with one token in length. If Thematizer still finds no match, then the parse returns the result 'Adverbial not in corpus,' and the marked theme remains unclassified. As the look-up table does not have every possible instantiation of marked themes, this output is then used to add the missing marked theme to the collection for future parses. As this example illustrates, the longer the marked theme, the greater the number of patterns Thematizer has to peruse. This is because the token length becomes the upper bound for the marked theme pattern to search for.

Once Thematizer does find a match in the look-up table, however, the corresponding marked theme type and semantic subclass are automatically extracted. Thematizer returns this information and saves the results for final output to the web interface.

While look-up tables can account for many marked themes and their varied realization, this can only partially resolve a marked theme's type and semantic subclass when certain adjuncts belong to multiple semantic classes. To illustrate this, sentences (2) – (5) all contain a marked theme starting with the preposition *in* highlighted in bold. Despite each sentence using the same preposition, each marked theme type is different: in (2), the prepositional phrase equates to a circumstantial theme on account of *book* being a locative or source; in (3), it is a modal theme through the set phrase *in fact*; in (4) a structural theme through *in spite of*; and in (5) a hypotactic theme due to *in order to*. In fact, the *in* in (5) is part of the infinitive phrase *in order to*, which is no prepositional phrase at all.

(2) **In the book**, Oliver discussed the outcome of the lawsuit.
(3) **In fact**, I think Dragonflight is probably going to be a universally good expansion.
(4) **In spite of that**, Akodon simulator has more proodont incisors.
(5) **In order to save his people**, he saw past the pharaoh's tricks.

The marked themes in (3) – (5) can be more readily identified and classified as they are conventionalized multiword expressions that cannot be modified. In other words, these phrases cannot undergo insertions through adjectives, such that *in amazing fact* or *in unlikely spite of*

would result in infelicitous expressions. Therefore, these marked themes were stored as such in the look-up tables and retrieved via the aforementioned search patterns. Only (2) presents an open-class adverbial, whereby the object of the preposition *in* can be realized with nearly any noun phrase. While the use of *the book* in (2) merits the marked theme's classification as circumstantial, it also indicates the semantic subclass of the expression as LOCATIVE. As soon as a different noun phrase is realized as the object of the preposition, the semantic class may change. For example, compared to (2), the marked theme in (6) below remains circumstantial on account of it being a prepositional phrase. However, its semantic subclass is TEMPORAL due to the temporal expression *2020* as the object of the preposition.

(6) **In 2020**, the shift caused many schools to scramble.

Thus, it is the both the semantic and syntactic information of the preposition that ultimately determines the semantic class of the overall marked theme.

Such semantic subclass disambiguation is only limited to adverbials that function either as subordinating adverbials or prepositions, and form the first set of exception cases to complete marked themes' classification parse. These adverbials, their semantic classes and sample realizations have been summarized in Table 4-13 for illustrative purposes but are not limited to these alone.

Adverbial	Possible Theme Type	Semantic Class Membership	Example
**as*	1) Hypotactic 2) Structural	1a) CAUSAL 1b) MANNER 2) ELABORATION APPOSITIVE	*1a) as little time is needed* *1b) as shown* *2) as a student*
to	1) Hypotactic 2) Circumstantial	1) INFINITIVAL 2) ANGLE	*1) to reach the store* *2) to some*
since	Hypotactic	1a) TEMPORAL 1b) CAUSAL	*1a) since 2010* *1b) since little time is needed*
in	Circumstantial	1a) TEMPORAL 1b) MANNER 1c) LOCATIVE	*1a) in 2010* *1b) in adjusting the score* *1c) in Germany*
from	Circumstantial	1a) TEMPORAL 1b) LOCATIVE	*1a) from 1967* *1b) from here*
between	Circumstantial	1a) TEMPORAL 1b) LOCATIVE	*1a) between 1750 and 1932* *1b) between here and there*
**for*	Circumstantial	1a) CAUSAL 1b) ANGLE 1c) TEMPORAL	*1a) for there was little time* *1b) for some* *1c) for years*
at	Circumstantial	1a) TEMPORAL 1b) LOCATIVE	*1a) at an earlier date* *1b) at the station*
by	Circumstantial	1a) TEMPORAL 1b) MANNER	*1a) by tomorrow* *1b) by using the term*
**on*	Circumstantial	1a) TEMPORAL 1b) LOCATIVE 1c) MANNER	*1a) on 16 July 1978* *1b) on the table* *1c) on beloved pastimes*

Table 4-13: Multiclass adverbials which need to be disambiguated through their syntactic parse for semantic classification. Here, the dependency parse, part of speech and/or .tag_ attribute as syntactic parameters for the adverbial's classification are required. Marked themes with an asterisk return multiple semantic classes in their classification results due to the ambiguity returned by certain dependency parses.

Adverbials as circumstantial themes require disambiguation most commonly as the object of the preposition often causes the semantic class to be either TEMPORAL or LOCATIVE. Since the dependency of the object of the preposition (POBJ) alone is insufficient in determining a circumstantial theme's semantic class, both part of speech and .tag_ attributes were used to delineate class membership further. The latter includes additional information on the word class and morphological information of the token in question.

To illustrate how these parameters are used for disambiguation, the marked themes in (7) – (9) with their semantic subclass in brackets have been provided.

(7) In 2010 [TEMPORALITY]
(8) In adjusting the score [MANNER]
(9) In Germany [LOCATIVE]

In each of these examples, the circumstantial theme starts with the preposition *in* and concludes with a temporal expression in (7), a nominalized verbal phrase in (8) and a noun phrase in (9). If Thematizer's syntactic parse returns a NUM part of speech for the object of the preposition (e.g., *2010*), then Thematizer assumes that a TEMPORAL circumstantial is present. If the marked theme contains both the .tag_ attribute VBG and the dependency PCOMP, which equate to that of a gerund and prepositional complement, respectively, then MANNER is assumed. Finally, if neither of these conditions is fulfilled, then Thematizer defaults the parse to LOCATIVE.

Dependency, part of speech and .tag_ attribute parses are performed on each marked theme adverbial case that requires disambiguation. While this approach accounts for nearly all ambiguous cases, a few cases of semantic classification had to remain generalized. These are marked with an asterisk in Table 4-13 above. If the marked theme to parse begins with *as*, *for* or *on*, then initial testing for semantic class takes place as outlined above. In the case of *as*, for example, if Thematizer determines the semantic class to be CAUSAL, then the combined output CAUSAL/MANNER is returned. Similarly, if Thematizer finds a positive case of the MANNER *on*, then its output is saved as LOCATIVE/MANNER. The reason for this generalization is because syntactic tests alone cannot account for the distinction between the specific semantic classes for those specific adverbials. Consider the following sets of text with the circumstantial themes in bold:

(10) **For students**, it may be difficult to find time. [ANGLE]
(11) **For consistency**, physicians were excluded. [CAUSAL]
(12) **As you would at home**, the heating should be turned on when needed. [MANNER]
(13) **As the exercise lacked real-world application**, few found interest in it. [CAUSAL]

In each of these cases, the syntactic parses of the marked themes are nearly identical. In (10) and (11), the noun phrases *students* and *consistency* both have the dependency POBJ and part of speech NOUN. In (12) and (13), a NSUBJ and congruent finite verb (*you would* and *the exercise lacked*, respectively) followed the adverbial *as* to form the subordinate clause. Hence, this syntactic information alone would be insufficient in determining the semantic class of the adverbial. Rather, the semantics of the marked theme is contextualized with respect to the semantics of the matrix clause. In other words, the semantic class of the circumstantial theme depends on the semanticity and propositional content that is expressed in the matrix clause. Without the matrix clause, the circumstantial theme *as you would at home* could be interpreted as *because you would at home* (CAUSALITY) or *in the same way you would at home* (MANNER). Through the propositional content expressed in *the heating should be turned on when needed* in (12), the semantic class of MANNER, not CAUSALITY, becomes evident.

As Thematizer is not equipped with semantic disambiguation capabilities on the basis of a text's overall semanticity, such limited cases are classified with multiclass membership in their output. Instead of outputting either ANGLE or CAUSAL for ambiguous cases of *for*, Thematizer outputs both together as ANGLE/ CAUSAL. If disambiguation can occur via syntactic testing alone, then Thematizer outputs the single semantic class that corresponded to the marked theme type as in (7) – (9).

In contrast to the marked themes discussed thus far, non-finite relative clauses and non-prepositional adverbials (e.g., *surreptitiously*) do not fall into a closed class of adjuncts saved in pre-defined look-up tables. This is because each can be realized in a nearly infinite number of ways. For example, the use of *when* as a marked theme is always a TEMPORAL hypotactic whose marked theme type and semantic class can be determined with the token *when* itself. For non-prepositional adverbials, every single adverb in the English language would have to be saved for such pattern-based matching, which is both infeasible and unnecessary. For these two exceptional marked theme classes, the structure-defining characteristic is, again, a syntactic one, which can be leveraged for marked theme classification. As shown in the following examples, non-finite relative clauses as in (14) are marked with the .tag_ VBN in their parse and can correspond to the semantic class of MANNER, TEMPORALITY or CAUSALITY; non-prepositional MANNER adverbials as in (15) have both an ADVMOD dependency and generally end in *-ly*.

(14) **Having little other choice**, the squirrel jumped into the river. [MANNER/CAUSAL]
(15) **Slowly**, the ink leaked from the pen. [MANNER]

Once pattern-based matching with the look-up table inevitably fails with these two cases since no matching token within the marked theme is found, Thematizer uses these syntactic conditions to test for their presence in the marked theme being parsed. If the VBN tag is found with the initial token of the non-finite relative clause, then it is defined as such. Similarly, if Thematizer identifies an ADVMOD dependency in an adverbial ending in *-ly*, then it classifies the marked theme as MANNER. As these cases are unique to non-finite relative clauses and non-prepositional MANNER adverbials, there is no need for disambiguation or further syntactic parsing.

Once the marked themes have been typified and semantically classified, Thematizer uses this information to create Spacy entities for the corresponding text spans. This allows for immediate retrieval of the marked theme type, semantic class, location in the text and textual realization in subsequent parsing steps and for final data visualization.

4.9.3. Summary of Marked Theme Parsing

As defined in the previous two sections, marked themes undergo two fundamental parsing passes. Once Thematizer has identified any marked themes within the sentence, their textual span is identified by means of their right edge index. This syntactic and indexical parameter allows Thematizer to account for the entirety of the marked theme phrase, whose length has a minimum of one in the case of single-word adverbials. If longer than one token, then the right edge index of the adverbial's head is required for Thematizer to locate the ending index of the marked theme span. Marked themes are extracted recursively until all have been processed.

Upon extraction, Thematizer leverages look-up tables which contain a pre-defined list of adjuncts for each marked theme type (circumstantial, structural, hypotactic and modal; projecting theme classification is based on syntactic and textual information alone). The extracted marked theme is compared against the pre-defined adjuncts in the look-up tables by

using token matching and the token length of the marked theme. The marked theme is then categorized according to the look-up table, in which it was found, i.e., if the marked theme was found in the circumstantial look-up table, then it was categorized as a circumstantial marked theme. The corresponding semantic class to which the marked theme belongs is also extracted from the look-up table upon a successful marked theme match. If a marked theme has multiple semantic classes, syntactic tests are performed for disambiguation based on the marked theme's realization.

After both extraction and classification have been performed, Thematizer saves the output in a JSON file, which is then used for data output and visualization in the web interface.

4.10 Thematic Progression Analysis

The third and final core thematic parsing step in Thematizer's thematic analysis is the identification of the thematic progression patterns across all sentences of the text. Figure 4-28 outlines the individual steps Thematizer progresses through, which are treated individually in the subsections to follow. Similar to how the themes and rhemes are extracted in the previous steps, the text is traversed sentence by sentence to determine the corresponding thematic progression pattern.

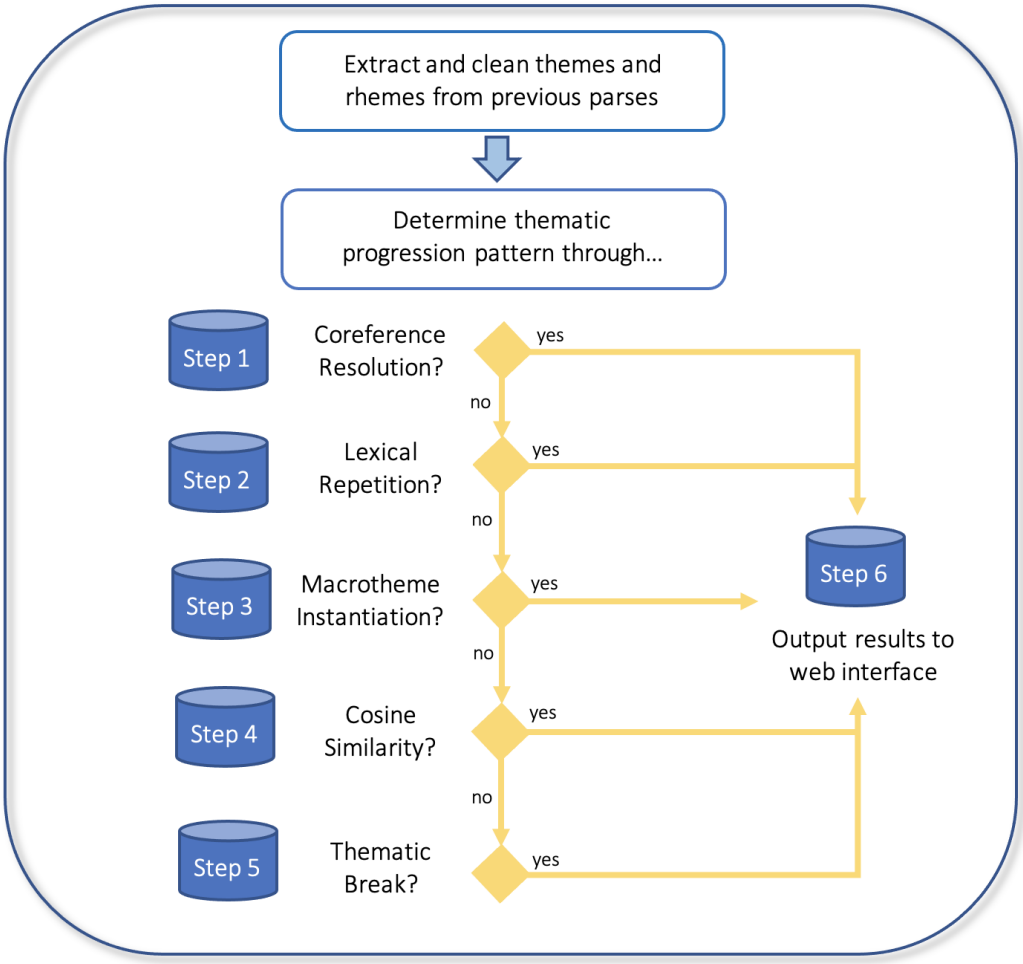


Figure 4-28: Parsing breakdown for determining thematic progression patterns across sentence clusters through coreference resolution, lexical repetition, macrotheme instantiation, cosine similarity and thematic breaks.

In general, the overall themes (i.e., marked and grammatical themes together) and rhemes for the current and previous sentences up to three sentences prior are extracted from the JSON

dictionary that was populated in the first two core parsing steps. Both themes and rhemes are cleaned again to remove cotextual noise in the form of extraneous punctuation and grammatical tokens in thematic progression's pre-processing steps. Here, grammatical tokens are lexemes such as prepositions, articles, conjunctions and auxiliary verbs. Once the themes and rhemes have been cleaned, Thematizer progresses through the syntactic and semantic tests illustrated in steps one to five in Figure 4-28. These tests then ultimately determine the thematic progression pattern present across sentence clusters. After identification, the progression pattern is added to the JSON dictionary for final data output and visualization.

Similar to the marked theme parse, this third and final thematic parse consists of a number of testing conditions, parameters and requirements. Since these are unique to each thematic progression test, they will be detailed in their own respective subsection in the following, starting with pre-processing the extracted themes and rhemes.

4.10.1 Theme and Rheme Pre-Processing

The first and one of the most important steps in determining the thematic progression pattern across sentence clusters is cleaning their respective themes and rhemes. Here, sentence cluster refers to a specific span of sentences. The current sentence *x* is always considered with respect to the previous sentence *x-n*, whereby $1 \leq n \leq 3$. Depending on which thematic progression test is executed, varying spans are required. However, each sentence cluster first examines the current and immediately preceding sentence (*x-1*). If no thematic progression is found across that span, then the current sentence is compared with the text from two sentences prior (*x-2*). This is repeated up to three sentences prior (*x-3*). Previous findings have indicated that the recall of previously established discourse topics becomes compromised three sentences after their realization (McCabe 1999: 176; Jalilifar 2009: 96). For that reason, only the themes and rhemes of the current and up to the previous three sentences are extracted at a time. The only exception to this is during coreference resolution tests, where only the immediately preceding sentence is compared against the current sentence. The reason for this is that the antecedent of an anaphor must be located within the previous theme or rheme for cohesion and coherence across sentences.

Compared to the text cleaning that occurs before theme and rheme index identification (cf. Chapter 4.7.1), this round of cleaning removes prepositions, articles, conjunctions, auxiliary verbs, punctuation and other stop words that Spacy has predefined. The reason for additional text cleaning before thematic progression analysis is to reduce cotextual noise that would cause subsequent tests to yield false positive results. Through the removal of punctuation and grammatical tokens, the core propositional content in each theme and rheme remains for the thematic progression tests that follow. To illustrate the pitfalls of performing thematic progression analysis without text cleaning, the following sentence clusters in (1) and (2) are provided. The sentence clusters in (1) have not been cleaned of stop or grammatical words while those in (2) have. The themes are in bold and marked in brackets.

(1) **The family** [T_1] is clearly lacking in Christian morals and strengths. **And rather than atoning for this, they** [$T_1 \rightarrow T_2$] go through hardship only to be rewarded with wealth in the form of marriage to wealthy men.
(2) **Family** [T_1] clearly lack Christian moral strength. **Atone this, they** [$T_1 \rightarrow T_2$] go hardship reward wealth form marriage wealthy man.

The thematic progression pattern in this sentence cluster is constant continuous progression since the thematic proform *they* in the second sentence refers to the thematic antecedent *family*

in the first sentence through coreference [$T_1 \rightarrow T_2$]. Without removing stop words before thematic parsing, however, Thematizer would return simple linear progression: the stop word *and* appears in both the rheme of the first sentence and the theme of the second sentence [$R_1 \rightarrow T_2$]. Neither coordinating conjunctions nor any grammatical lexis can instantiate thematic progression through their repetition as they lack propositional content, i.e., they serve a grammatical function alone. Hence, without removing non-propositional content, Thematizer would invariably return incorrect thematic progression patterns and incorrect progression instantiators. This problem is then circumvented by cleaning the text of such stop words and grammatical lexis.

In addition to removing cotextual noise, tokens are lemmatized during pre-processing to reduce them to their base form, i.e., the dictionary form without any morphological changes to the token. This can be seen when comparing *atoning* in its inflected form in (1) with the lemmatized *atone* in (2). Without lemmatization, Thematizer would fail to recognize reinstation of lexis as another word class or part of speech. For example, if the theme in one sentence were *women* but then realized as *woman* in the theme of the next sentence, Thematizer would assume these tokens to be unrelated. Lemmatization thus equalizes the tokens across the sentences being compared, which is critical for thematic progression via lexical repetition and for identifying the tokens responsible for instantiating thematic progression.

A final point to mention about text cleaning is that personal and demonstrative pronouns were explicitly ignored, i.e., not removed. While these two classes of pronouns are conventionally included in stop word lists, their removal would cause thematic progression analyses to fail. In blogs and lyrics in particular, personal pronouns are a frequent occurrence since the recipient of the text is often addressed directly. Were the first-person pronouns removed from the following sentence cluster:

(3) **I** cannot recommend this recipe enough. **I** have linked it at the bottom of the page for you.

the bold themes as grammatical subjects would be deleted from the data and the thematic parse would return sentences without any grammatical themes. With the erroneous removal of the themes entirely, the parser would yield rhematic progression via coreference resolution through *recipe → it* across the rhemes instead of the correct pattern of constant continuous progression through repeated use of *I*. In the same vein, demonstrative pronouns, either as endophoric or exophoric deixis, would fall victim to this problem if *this* or *these* were the sole theme in a sentence. Their removal would equate to the removal of the grammatical subject, which would automatically force rhematic progression or a thematic break due to a sentence without a theme. As such, coreferential terms remain intact after cleaning. Once the theme and rheme spans are cleaned, they are passed on to the respective thematic progression tests.

4.10.2 Thematic Progression via Coreference Resolution

The first thematic progression test that Thematizer performs is coreference resolution. This determines whether demonstrative pronouns and adjectives (*this, these, those*) or proforms (*s/he, it, they, their*) anaphorically refer to an antecedent in the immediately previous sentence. The tool Coreferee was used to determine coreference chains across sentence clusters, which are then resolved indexically during this test (cf. Chapter 4.6). While cataphoric references may also be present, they are programmatically resolved as if they are anaphora since Coreferee saves them as such. For simplicity's sake, therefore, all cases of coreference resolution are subsumed under and referred to as anaphoric resolution.

Functionally, Coreferee identifies the indexical location of anaphora-antecedent chains throughout a text using a rule-based approach. These chains are stored as so-called chain objects, which are comprised of mention objects that hold the anaphor, antecedent and their token indices. Akin to accessing theme and rheme spans via their index in the first thematic parse, Thematizer can access relevant mention objects to resolve instances of coreference in a text and retrieve the tokens in the anaphora chains for processing. Thematizer then saves all cases of anaphora from the text for reference and access when progressing through the thematic progression tests sentence by sentence.

When Thematizer attempts to identify thematic progression across a sentence cluster via coreference resolution, the indices returned by the Coreferee parse are compared to the indices of the current sentence cluster. While it may seem trivial at first, ensuring that the correct indices are accessed with respect to the sentence cluster under consideration is critical.

To help understand the importance of ensuring updated coreference indices vis-à-vis sentence indices, consider the following text excerpt with a coreference chain occurring between sentences (1b) and (1c) only. The anaphor-antecedent pair is marked in bold and with the token's index in the subscript.

(1a) Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing.
(1b) Many different classes of machine-learning **algorithms**$_{32}$ have been applied to natural-language-processing tasks.
(1c) **These**$_{40}$ take as input a large set of "features" that are generated from the input data.

Since not every sentence has (anaphoric) coreference to resolve, as is the case in (1a) and (1b), Thematizer must determine whether the coreference chains identified by Coreferee fall within the sentence spans currently being analyzed. This is first accomplished by determining the index spans of the sentence pairs in question. For (1a), the sentence span is 0–26, where each token (and punctuation mark) is given an index. The index span for (1b) is 27–39 and finally 40–58 for (1c). During parsing, Coreferee will have identified a coreference chain between *these* in (1c) and *algorithms* in (1b), whose index pair is 40:32. Index 40 then becomes the relevant index that Thematizer needs to search for within sentence spans.

When Thematizer reaches the first sentence, it compares its span of 0–26 with the anaphor's index of 40. As 40 does not fall within the first sentence's index span, Thematizer concludes that no coreference resolution is necessary in the first sentence.[22] In such cases, the coreference resolution test fails as intended and Thematizer moves on to the next thematic progression test. The same is true once Thematizer progresses to the second sentence: again, the anaphor's index of 40 is not within the second sentence's index span of 27–39. At the third sentence, however, the test returns positive due to the index of *these*, 40, being within the sentence span 40–58.

With this result, Thematizer initially marks the sentence pair as developed via coreference resolution. It is at this point that the index of the antecedent *algorithms*, namely 32 in (1b), becomes important. Just as sentences each have their own index span, themes and rhemes also span a given portion of the self-same sentence. Then, depending on where the antecedent's index is located within the sentence, the parser can determine the corresponding thematic

[22] The first sentence of the text is always ignored since there is no previous sentence before the absolute first sentence of a text upon which to base thematic progression. In text excerpt (1), therefore, it should be assumed that there are sentences that actually appear beforehand.

progression pattern. Here, Thematizer also ensures that the textual realization of the antecedent was found in the span, i.e., that *algorithms* as a token was realized in the previous sentence. The index span for the theme of (1b) is from 27 to 32, the rheme from 33 to 39 (cf. Figure 4-29 in the following). The antecedent's index 32 lies within the theme span, which would thereby yield constant continuous progression between (1b) and (1c) since the anaphor *these* from (1c) is also within the theme.

Thematicity	**THEME**						**RHEME**
Sentence (1b)	*Many*	*different*	*classes*	*of*	*machine-learning*	*algorithms*	*...*
Token's Index	27	28	29	30	31	32	33
Anaphor's Index	-	-	-	-	-	40	-

Thematicity	**THEME**	**RHEME**
Sentence (1c)	*These*	*...*
Token's Index	40	41
Antecedent's Index	32	-

THEME → THEME
Constant Continuous
Progression through
algorithms → *These*

Figure 4-29: Coreference resolution via Coreferee's anaphor-antecedent indices as coreference chain markers in text. These indices are then used to determine where the antecedent was realized in the previous sentence for thematic progression classification.

If an anaphor has multiple antecedents that, in turn, span multiple sentences, Thematizer must cycle through the coreference chain indices until an index has been reached that is lower than the index of the anaphor in the current sentence. For example, a sentence might have an anaphor with index 42, which forms a coreference chain with tokens at indices 4, 10, 28, 36 and 52. Assuming only index 36 occurred within the immediately preceding sentence of index 42, that is the index that Thematizer would need in order to resolve coreference and determine the corresponding thematic progression pattern. Thematizer would therefore cycle through the list of antecedent indices until index 36 is within the span of the previous sentence and the anaphor's index 42 is within the sentence span of the current sentence being analyzed.

The reason why it is necessary to update the antecedent's index while parsing is alluded to in the list of indices for this previous example. The anaphor has an index of 42, which forms a chain with five potential antecedents within the text. If Thematizer tried to access the antecedent's index beyond the span of the previous or current sentence, it would throw a so-called Index Out of Bounds error and cause the program to crash. This is because Thematizer would try to access a token and its index that do not fall within the sentence spans currently being analyzed. In other words, Thematizer would attempt to access a token that Coreferee – and Python by extension – assumes to be non-existent. Automatically updating the referent's index before, during and after a sentence parse prevents this error from occurring while maintaining the correct coreference chain pair for the current sentence. It also ensures that the relevant anaphor-antecedent pairs are considered for identification of the sentence cluster's thematic progression pattern.

A final note concerns the demonstrative pronouns *this* and *that*. These two cases can either point back to single noun phrases or to entire phrases as in:

(2) It is computationally expensive to analyze hundreds of thousands of texts. **This/That** is an important step, nonetheless.

In this example, the demonstrative pronouns can either refer to a single sentence constituent in the previous sentence or to the sentence in its entirety. Since Coreferee does not account for this kind of coreference resolution, Thematizer qualifies such cases as linear progression as a default.

Once an anaphor and its antecedent have been correctly resolved, the thematic progression pattern and relevant information between the sentences in question are saved. In addition to the thematic progression pattern itself, Thematizer also stores whether the anaphor's antecedent stemmed from the previous theme or rheme, the means of progression as coreference resolution, and the antecedent token as text found to match the anaphor. This information is appended to a final output list, which is later used when extracting the analytical results for data visualization. If no coreference needs to be resolved across sentences clusters, Thematizer skips this test and moves on to lexical repetition as a means for thematic progression.

4.10.3 Thematic Progression via Lexical Repetition

The exact or partial repetition of a lexeme or phrase across sentence clusters constitutes the second thematic progression test that Thematizer performs. As opposed to the coreference resolution test, where only the immediately preceding sentence is used for comparison, lexical repetition tests compare the current sentence with up to the previous three sentences.

For parsing, a sentence gap parameter was defined to determine the number of sentences Thematizer must look back for lexical repetition. This always starts with a gap of one, which indicates that the themes and rhemes from the current sentence are compared with the themes and rhemes from the immediately preceding sentence. Providing a sentence gap firstly simplifies the code by reducing redundant calls to the lexical repetition method. Secondly, it allows specific sentence ranges to be set beforehand when the parsing function is called. Finally, it ensures that Index Out of Bounds errors do not arise since the sentence gap is always provided with respect to the present sentence.

The location of lexically repeated themes and rhemes across the sentence clusters being compared ultimately determines the thematic progression pattern. Should lexemes be repeated across both themes, then constant continuous progression is present. Repetition from a previous rheme to the current theme constitutes linear progression. These progression patterns are then considered gapped, i.e., gapped continuous or gapped linear progression, if lexical repetition is found two or three sentences prior.

In lexical repetition tests, Thematizer considers two realizational patterns: either compound noun phrases or singular lexical repetition. Lexical repetition can take the form of identical repetition, such as repeating *car* across sentence clusters. Otherwise, lexical repetition can occur as a derived or related form of a given word. For example, Thematizer would consider the recurrence of the lexemes *experiment* and *experimental* as lexical repetition since *experiment* is repeated as an adjective. In order to establish uniformity across the lexemes to be compared, the noun chunks from the themes and rhemes are lemmatized. Without doing this, Thematizer

would not be able to distinguish correctly between singular and plural or words that belong to the same word family albeit as a different part of speech.

Thematizer considers the lemmatized thematic and rhematic noun chunks first in terms of potential compound noun phrases. Consider the following sentence pair with constant continuous progression on account of lexical repetition within the bold themes:

(1) **Natural language processing** has advanced. In contrast to previous approaches to **natural language processing**, modern ones make use of machine learning.

Programmatically, Thematizer is provided with the lemmatized noun chunk *natural language process* from the first theme and the list of noun chunks *previous approach, natural language process, modern one* from the second theme. Thematizer then cycles through the list of compound nouns of the current sentence to search for a match with the noun chunks of the themes and rhemes from the previous sentence. A step-by-step breakdown of this search is given in Table 4-14. Lexeme pattern matching for compound nouns is done with the Python IN operator, which returns a True condition if exact textual realizations are present in a given text string or list of strings.

	Noun Chunk from Current Sentence	Noun Chunk from Previous Sentence	Repetition Found?
Lexeme Matching Pass 1	*previous approach*	*natural language process*	False
Lexeme Matching Pass 2	*natural language process*	*natural language process*	True

Table 4-14: Breakdown of noun chunk matching parse from the current theme compared to previous themes and rhemes. This repeats either until a match is found or all noun chunks have failed to return a match.

Compound lexical repetition is performed first because of the greater degree of complexity when compared to single lexical repetition. If single lexemes were compared first, then Thematizer would have still found lexical repetition in (1); however, only *natural* would have been returned as the repeated lexeme since it is the first item in the compound noun. Therefore, while the correct thematic progression pattern would have been identified in this case, that is not guaranteed. Further, this parse would have resulted in partial extraction of the repeated element only. Therefore, the output would have been incomplete and only partially correct.

If no compound nouns are present across sentence clusters, then Thematizer searches for lexical repetition with single lexemes. Functionally, lexeme matching follows the same approach as outlined in Table 4-14 above, albeit with individual lexemes. Instead of only noun chunks as search terms, themes and rhemes may contain adverbs, adjectives, nouns and verbs after having been cleaned of grammatical tokens. These are also lemmatized to create uniform tokens. The resulting lemmatized tokens then populate a list for subsequent comparison. Here, Python's IN operator is used again to determine whether an instance of a single lexeme is present in the list of themes and rhemes. Where the repeated lexeme is found, either in a previous theme or rheme, determines the resulting thematic progression pattern.

In initial development of Thematizer, the regular expression method RE.SEARCH was used to find matches of repeated single lexemes. If the lexeme *test* was provided as the search term, the RE.SEARCH method would only return a positive match when an exact match was found. Therefore, if the current theme *test* was compared against the previous theme *testing*, no successful match would be returned since these lexemes are not identical. As lexemes realized

as a different part of speech are considered lexical repetition, however, this approach proved insufficient. This problem was then circumvented through the IN operator, which would return a positive match since *test* partially constitutes the lexeme *testing*. While the IN operator affords greater flexibility in string matching, it can lead to false positives. For example, a positive match could be found between the tokens *man* and *German* since the former comprises part of the latter. Despite this potential for false positives, the IN operator allows for accurately identifying derived lexical repetition at greater frequency.

Any successful cases of lexical repetition are stored as a bundle of relevant thematic progression constituents. This contains the thematic progression pattern found, the sentence where repetition was found, whether compound or single lexeme repetition occurred and the actual token(s) repeated. Should no lexical repetition be found within the three previous sentences, then Thematizer considers this test a fail. A failure in this test then prompts Thematizer to progress on to the next thematic progression test, macrotheme instantiation.

4.10.4 Thematic Progression via Macrotheme Instantiation

Macrothemes are discourse topics that have achieved statistical significance on account of their n-gram frequency distribution in the text (cf. Chapter 4.6). Statistical significance is calculated via Latent Dirichlet Allocation, which is a built-in function that the library Gensim offers and is implemented in Thematizer's parsing functionality. Gensim was chosen specifically for its ease of implementation, multithreading capabilities and its memory independence, i.e., Gensim can process vast amounts of text input larger than the available RAM via streamed data processing (Řehůřek & Sojka 2010).

While multiple documents are typically used as input for Gensim to determine statistically relevant topics, Thematizer instead considers each sentence as a single document for macrotheme identification. Upon beginning the parse for macrotheme instantiation, Thematizer extracts the noun phrases from the entire text to be analyzed. Since noun phrases constitute the text's discourse topics, these are fed into the Gensim model for calculation of statistical significance, assuming the noun phrase occurred at least twice within the text. Minimum frequency is a parameter that can be set to tune the output of a text's potential discourse topics. As Thematizer users can input a text of any length, ensuring this parameter enabled suitable coverage of discourse topics without overgeneralization was pivotal. Increasing this parameter to higher minimum frequencies would result in macrothemes remaining overlooked; a minimum frequency of one, however, would overinflate the list of discourse topics that Gensim output. In light of these conditions, a minimum frequency of two, while low, proved to offer appropriate coverage without compromising the validity of the other thematic progression tests.

Once fed into Gensim, the relevant noun phrases are converted to their bag-of-words representation of data. This vectorizes all noun phrase tokens, such that each unique token is appended to a list that represents the unique vocabulary and dimensions of the entire input text. For instance, if a text consists of 20 unique tokens, the vectorized token list has 20 dimensions. This list is then complemented by the frequency of the tokens appearing within the individual noun phrases from each sentence. This frequency distribution for a sentence is then the vectorized representation of the noun chunks from a specific sentence. Taken together, a multidimensional matrix of total unique tokens and the frequency of tokens comprising the input noun chunks per sentence constitute the numerical breakdown of an entire text. This numerical data is then required for Gensim to compute which noun phrases are of particular relevance in a given text.

Gensim ultimately determines discourse topics by using the vectorized text and the so-called topic coherence of each sentence's noun phrases. Topic coherence, also known as a topic's strength, measures how much a noun phrase statistically contributes to a text's overall discourse message and falls between zero (irrelevant) and one (most relevant). Due to this range, topic coherence is also considered the probability that a given noun phrase will appear with a certain topic. For example, the noun phrases *matrix, gradient, solution* and *gradience* are commonly associated with the topic *machine learning* and may have a topic coherence of 0.54, 0.39, 0.32 and 0.24, respectively. In a text on machine learning, the topic coherence then represents the probability of these noun phrases' use in the text. Discourse topics and topic coherence thus vary depending on the number of texts provided for topic identification, the number of topics to be identified and the length of the texts. For statistical representativity, the longer the text and/or the greater the number of texts, the more reliably the results can be interpreted. Since Thematizer uses sentence number as the number of documents, the longer the text that the user provides, the greater the likelihood that reliable and statistically representative topics can be identified.

Using topic coherence as a statistical measure to weed out less relevant topics from a text, Thematizer sets a lower bound of at least 0.1 topic coherence. A topic coherence of 0.1 for the macrothemes to be identified allows Thematizer to ignore low-frequency, less discourse-relevant topics. A pre-defined number of topics set to four simultaneously ensures sufficient topics to be extracted as a discourse-relevant topic from the text. These specific parametric values were decided upon during development and testing of Thematizer.

Once the macrothemes have been ascertained, these are saved as a list for Thematizer to cycle through during macrotheme instantiation testing. In this test, Thematizer first extracts the theme from the current sentence, which has been cleaned and lemmatized. This is compared against the list of the individual macrothemes using the regular expression method RE.SEARCH. This method was employed for exact matching to avoid partial matches. If an exact match is found, then the parser marks the theme as a macrotheme and appended to the final thematic progression output list.

Testing for macrotheme instantiation after lexical repetition was a deliberate choice for the order of thematic progression testing. A theme can only be considered a macrotheme if a) Gensim identifies it as a statistically significant discourse topic, and b) the thematic noun phrase does not appear in any of the previous three sentences before the current sentence. If the same lexemic phrase is found across a sentence cluster within the three-sentence span, then it may not qualify as a macrotheme. Instead, this qualifies as thematic progression via lexical repetition.

Thematic progression via macrotheme instantiation represents the final parsing test that Thematizer conducts either via indexical or string-based pattern matching. At this point of testing, the majority of thematic progression cases will have been accounted for, as the present research has found. The next round of thematic progression testing, cosine similarity, presents a shift from syntactic to semantic testing that attempts to resolve thematic development across sentence clusters.

4.10.5 Thematic Progression via Cosine Similarity

The final test for determining thematic progression is conducted by comparing the semantic similarity of the theme and rheme from the current sentence to those from the previous three sentences. Such semantic tests are conducted to determine whether thematic progression occurred via lexical entailment across sentence clusters. Potential lexical entailment is

calculated via so-called cosine similarity by using Spacy's built-in similarity method. This uses word embeddings, i.e., numerical representations, of word meanings that are constructed using algorithms such as word2vec or Multi-Sense Skip-Gram (Honnibal et al. 2020c). The resulting similarity value is then an average of the token vectors being compared in a given context. A value of 1.0 means that the compared tokens are identical; -1.0 means that they are entirely unrelated.

Depending on the similarity value returned, the parser is then able to determine the thematic progression pattern present, if at all. Because there is no gold standard or otherwise conventional threshold that similarity values must achieve to determine thematic progression patterns, the varying thresholds for a given thematic pattern first needed to be defined. For this, the thematic progression patterns for each sentence cluster in the 150 test texts were manually determined. Second, the cosine similarity between the themes and rhemes of the identified thematic progression pattern was calculated and tallied in a separate database. For instance, assuming gapped linear progression was identified between two sentences, the cosine similarity value between the theme of the current sentence and the rheme of two sentences prior was calculated (e.g., 0.53). This value was then stored in a database along with the thematic pattern gapped linear. Once all similarity values had been calculated and stored, the minimum and maximum threshold of the corresponding thematic pattern was determined. This resulted in the thresholds summarized in Table 4-15:

Progression Pattern	Threshold	Sentence Position Relative to Current Sentence *n*
Constant Continuous	$0.28 < x < 0.66$	Sentence *n-1*
Linear	$0.30 < x < 0.64$	Sentence *n-1*
Rhematic	$0.68 < x \leq 1.00$	Sentence *n-1*
Gapped Constant Continuous	$0.25 < x < 0.61$	Sentence *n-2*
Gapped Linear	$0.19 < x < 0.62$	Sentence *n-2*
Triple Gapped Constant Continuous	$0.15 < x < 0.62$	Sentence *n-3*
Triple Gapped Linear	$0.20 < x < 0.61$	Sentence *n-3*

Table 4-15: Upper and lower bounds for similarity values x *used in determining the corresponding thematic progression pattern. Where the cosine similarity value fell within the threshold ultimately yielded the thematic progression pattern between sentence clusters.*

The lower value in the threshold column represents the minimum similarity value, the higher value then the maximum similarity value for each progression pattern. The similarity value ranges for continuous, linear and rhematic progression patterns were 0.40, 0.39 and 0.32, respectively. A narrower similarity value range for rhematic progression is advantageous as it is considerably rarer than its continuous and linear counterparts. A smaller range for rhematic progression and a larger range for continuous and linear progression thereby allows Thematizer to more readily account for typical distribution patterns in thematic realization patterns. Conversely, similarity value ranges for gapped progression patterns were between 0.36 and 0.47.

It is important to note that triple gapped progression is denoted as such here for nomenclature purposes only. In the output, if gapped progression is found between the current sentence and three sentences prior, it is marked as gapped progression without the *Triple* modifier. In

Thematizer, a programmatic distinction had to be made to accommodate for the varying threshold values and progression pattern.

These threshold values constitute one of two conditions to be fulfilled when determining which progression pattern is present across sentence clusters: Firstly, the similarity value must fall within the threshold as defined in the table above. This then initially marks the corresponding progression pattern as a potential candidate. In order to finally confirm the progression pattern, however, its similarity value must secondly be greater than the competing progression pattern's value. Each progression pattern can be seen as a pair of thematic and rhematic progression: constant continuous progression as a thematic pattern and linear progression as the competing rhematic pair; gapped constant continuous progression as thematic and gapped linear progression as the competing rhematic pair; finally, triple gapped constant continuous progression as thematic and triple gapped linear progression as the competing rhematic pair.

Therefore, when the parser tests for constant continuous progression, it first calculates the similarity value between the two concomitant themes. Then, the similarity value between the previous rheme and the current theme is calculated for linear progression. If the tests yielded a value of 0.60 for constant continuous and 0.33 for linear progression, for instance, the similarity value for constant progression would not only fall within the constant continuous threshold ($0.28 < x < 0.66$) but also be larger than that of linear progression ($0.60 > 0.33$). As such, Thematizer would mark the progression across the respective sentence clusters as constant continuous. Conversely, if similarity values of 0.60 for constant continuous and 0.63 for linear progression were returned, linear progression would be returned since its similarity value fell within the linear threshold and was larger than the constant continuous value.

With these conditions in mind, the exact process behind Thematizer's cosine similarity parse for thematic progression pattern identification can be explained. The parser first ensures that the second sentence has been reached so that it can be compared to the previous sentence. While thematic elements are reduced to noun, adverbial and adjectival phrases alone, rhematic elements only have their articles, prepositions and conjunctions removed. The reason for this is due to the constituents a rheme can have. It is entirely possible for a sentence to be without a theme, such as in imperatives or subject-less fragments. However, every sentence must have a rheme, which, at its bare minimum, can simply be the finite verb, e.g., *Go!* Therefore, any sentence constituents that contribute to the propositional content of the sentence are retained when cleaning the rheme.

With the cleaned themes and rhemes, testing for thematic progression via cosine similarity follows the parsing structure of exception cases first, then basic thematic progression tests, followed by gapped progression and finally triple gapped progression tests. The first exception case is sentences that have no theme. Here, either rhematic progression or a thematic break must be present. Rhematic progression is present if the similarity value between both the previous rheme and the current rheme falls within the threshold for rhematic progression. Should the similarity value not fall within the threshold, then Thematizer marks the progression pattern as a thematic break. A break in progression thereby indicates that all potential forms of thematic progression failed, which is common when shifting to new rhetorical sections of a text or when no connecting elements can be found that merit progression across sentence clusters.

The second exception case is the use of the demonstrative pronouns and adjectives *this* and *that* which refer to the partial or entire content of a previous sentence, as shown in bold in (1) and (2). With these, the parser defaults to linear progression to indicate that *this* or *that* functions as a theme that builds off the propositional content presented in a previous sentence.

(1) It is computationally expensive to analyze hundreds of thousands of texts. **This/That/This approach/That approach** is an important step, nonetheless.
(2) Text analysis is a central to text linguistics. **This/That/This task/That task** involves consideration of a text's structural or developmental breakdown, amongst other things.

Demonstrative pronouns and adjectives realized as the theme in the current sentence can instantiate constant continuous progression by building off a thematic referent from the previous sentence as in (2). Since Coreferee does not resolve these forms of coreference, Thematizer attempts to resolve them via cosine similarity. Testing in the development of Thematizer proved, however, that similarity values in these cases always favored linear progression even when constant continuous was present as in (2). For that reason, Thematizer defaults to linear progression with demonstrative adjectives and pronouns to reduce processing complexity and time, albeit at the potential expense of parsing accuracy.

The final group of exception cases in cosine similarity testing is clefts and existentials. Neither of these structures fulfills a coreferential function. Instead, the thematic dummy-*it* in the cleft and the *there* in the existential function to highlight or emphasize the information that follows in the rheme; the non-referential themes establish a rhetorical basis for the corresponding rheme. In turn, the discourse message is developed via the rheme only so that any progression instantiated via the theme can be ignored. As such, clefts and existentials automatically default to rhematic progression.

If there are no exception cases present in the text to parse, then Thematizer first calculates the similarity values between the theme and rheme of the current and immediately preceding sentence. If no thematic progression pattern is identified here, then tests for gapped progression ensue. Similarity values are calculated again, this time between the current sentence and two (gapped) or three (triple gapped) sentences prior. As is the case with testing across all sentence clusters, the calculated similarity value must fall within the respective pattern's threshold and be larger than the competing, i.e., thematic vs. rhematic, pattern.

Thematic progression via cosine similarity concludes with testing for a thematic break. This pattern is achieved when all other previous tests for progression have failed. In other words, a thematic break indicates that progression across sentence clusters could not be resolved via coreference resolution, lexical repetition, macrotheme instantiation or cosine similarity. A thematic break thereby represents a final failure case for thematic progression.

Once Thematizer identifies the progression pattern evident across sentence clusters, it returns the following information to be added to the thematic progression results collection: the progression pattern identified; whether progression built off a previous theme or rheme; means of progression, e.g., via paraphrase or existential; and finally, the previous theme or rheme that instantiated progression as a text string. After thematic progression analyses have concluded and their results have been added to the collection of the text's thematic progression breakdown, Thematizer concatenates this to the JSON file containing all thematic analysis results.

4.10.6 Summary of Thematic Progression Parsing Task

In Thematizer's final parsing task, the parser determines whether thematic progression was instantiated as constant continuous, linear, rhematic or gapped progression by means of five tests. First, coreference resolution is tested by using coreference chains that Coreferee identifies. The location of an anaphor's antecedent in the theme or rheme of the immediately preceding sentence determines constant or linear progression. Partial or exact lexical repetition of compound nouns or individual lexemes across themes and rhemes constitutes the second test for thematic progression. Third, macrotheme instantiation merits thematic progression through the use of statistically significant discourse topics identified by Gensim. Fourth, cosine similarity tests for lexical entailment calculate the semantic similarity between themes and rhemes to determine the thematic progression pattern. Fifthly and finally, Thematizer returns a thematic break if none of the previous tests finds evidence of thematic progression.

Upon completion of thematic progression tests, the entirety of a text's thematic analysis is concluded, whose results are then passed on to the Dash framework. This takes the analytical results as input and produces visualized output in the form of highlighted text, frequency figures and summarized tabular data. The final output that users can interact with via the web interface is thereby a reflection of all syntactic and semantic analyses outlined up to this point in the dissertation.

4.11 Thematizer Web Interface via Dash

In this section, the user interface produced via Dash and its accompanying library Plotly is presented in their entirety. The web-based program is broken down into the start screen and four results tabs. Once the user has entered a text on the start screen, the analytical results are presented according to their respective parses.

The first results tab reproduces the text input with highlights to indicate the marked theme(s), grammatical theme and rheme of each sentence (cf. Figure 4-30 and Figure 4-31). The realization of each marked theme and where it can be found in the text is also presented purely as a visual, descriptive breakdown of their emergence in the text.

The second results tab summarizes the frequency of the text's marked themes and their semantic class as a bar chart (cf. Figure 4-32). How these marked themes are realized within the text is also summarized in a corresponding table. This table extracts the marked theme span as realized in the text and the sentence in which it is used.

The third results tab summarizes the thematic progression patterns for each sentence of the text (cf. Figure 4-33). Frequency distributions for each thematic progression pattern and means of progression are illustrated in a bar chart. The input text is reproduced here with highlights again, however with respect to thematic progression only.

Finally, the fourth results tab tallies the text input's frequencies of thematic progression patterns, marked themes and means of progression in comparison to five other text types (cf. Figure 4-34). The text types for comparison were the same used for the testing and development of Thematizer: Wikipedia articles, L1 and L2 university texts, blog articles, and lyrics. All data here is summarized in three separate figures, which are dynamically populated during use, as is the case with the results in the previous tabs.

In the following, the web interface and the output from sample analytical results are described. The discussions do not explain the code required to produce the output since it primarily took the form of extracting the requisite data from the compiled data output JSON file using the Panda library alone. Instead, the explanations below simply outline the constituent makeup of the web interface, its functionality and the relevant information required from the thematic analysis parses. Screenshots of the respective web interface components have been provided and will be used as reference points for the explanations.

4.11.1 The Start Screen

When the user accesses Thematizer, the start screen in Figure 4-30 is loaded, where a welcome text first greets the user at the top. There, a general introduction to the purpose and functionality of the tool is presented. Minimal text was aimed for so as to avoid overwhelming the user with instructions and to establish a more minimalistic layout. As indicated both in the welcome text and the instructions within the text field, a sample text is automatically loaded so users can interact with the program without uploading anything first. This affords them the ability to explore the various analyses and visualizations before inputting their own text(s). The text provided was written by the present author to explain the purpose of Thematizer and to exemplify all marked theme types and thematic progression patterns.

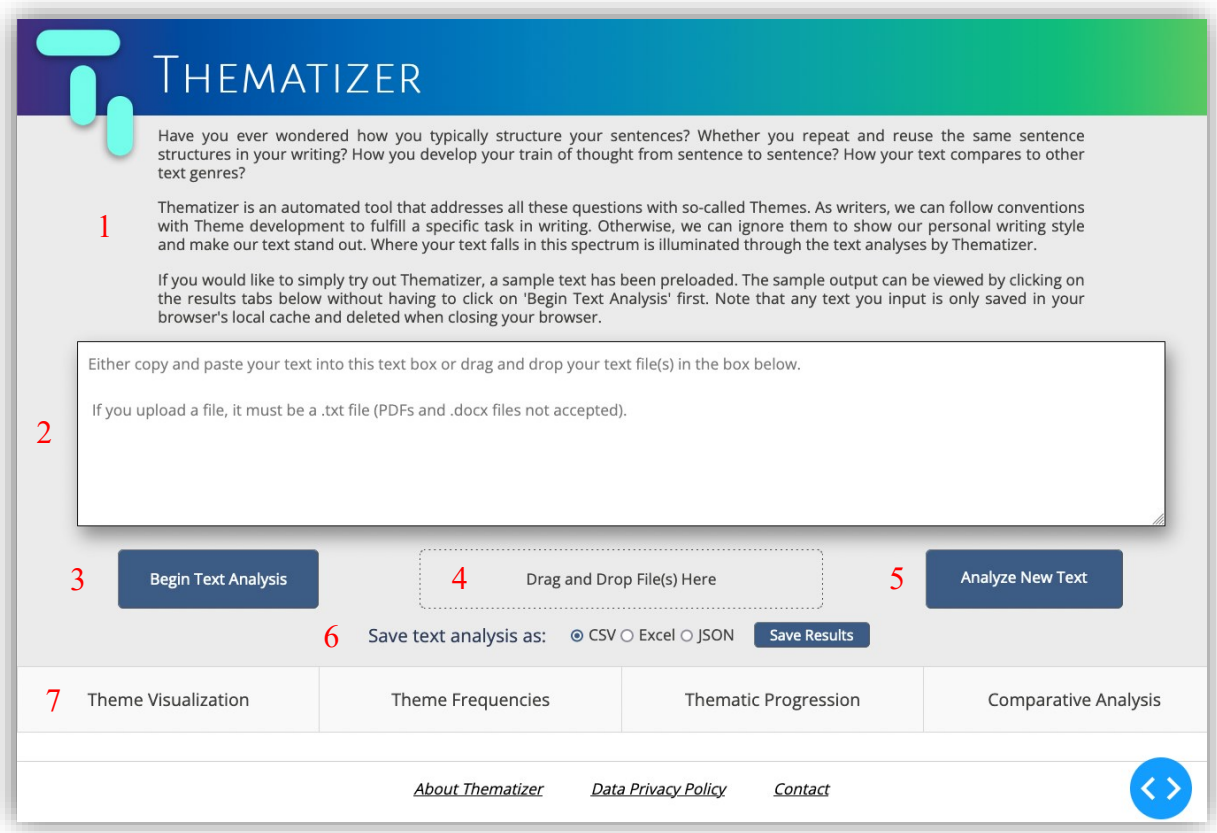


Figure 4-30: Start page that is loaded when the user first accesses Thematizer via the web interface, which includes an introductory text, a text field for input and the various buttons for initializing and saving the thematic analysis.

Beneath the welcome text (cf. [1] in Figure 4-30), the user can find the textbox where text can be input for processing (cf. [2]). Alternatively, users can drag and drop single or multiple text files into the corresponding box (cf. [4]). Requiring text files with .txt file type only was for processing reasons: programmatically, files of any text type are possible, such as .doc(x) or PDF; however, as the formatting according to the data type may vary widely and thus cause errors in how the text is pre-processed, supplying .txt files helped to obviate these errors.

Once the user has entered or uploaded a text in the text field, they can click on the 'Begin Text Analysis' button to initiate processing (cf. [3]). Text analysis in progress is indicated by an animated loading icon upon click the button. Once analysis is complete, the text field disappears, and the result tabs scroll up to indicate that the results can be viewed. The user then has the option to analyze a new text by clicking on the button 'Analyze New Text' (cf. [5]) or they can save the text analysis as a CSV, Excel or JSON file via the 'Save Results' button (cf. [6]). These file types were provided to ease incorporation of the processed data into the user's own corpora or for use in other linguistic analyses. Once the text has been analyzed, the user can scroll down to the different tabs, where specific results are visualized depending on their purpose (cf. [7]).

Finally, separate pages for information about the development of Thematizer, its data privacy policy and contact are provided at the bottom of each page. The 'About Thematizer' page summarizes the developmental background and impetus behind the program. The 'Data Privacy Policy' explicitly outlines how the user's data is neither saved nor stored; results are only saved in the browser's cache but then deleted once the user closes the browser entirely. Finally, should users be interested in contacting the present author and developer of Thematizer, they can do so via the 'Contact' link.

4.11.2 Results Tab 1: Theme Visualization

In this first results tab, the thematic constituent analysis is presented via a marked theme realization chart and individual highlights within the user's text). At the top of this and each tab, the user can choose to show or hide generalized explanations of the results summarized in the respect tab (cf. [1] in Figure 4-31 in the following). The text that appears was written for a general audience without previous knowledge of thematic theory. Ultimately, it outlines what information is being presented in the tab, what the linguistic terms used in each tab mean and how users can use these to understand their results.

Beneath the results explanation button, a drop-down menu is provided for the user to cycle through the results from their respective texts if more than one text was uploaded (cf. [2]). The individual texts populate the drop-down menu after processing and use the same name as the actual text file uploaded. This allows the user to easily identify and reference the specific analyzed text.

Selecting a document name from the drop-down produces the corresponding text with its resulting highlights and a chart that visualizes where marked themes were used in the text (cf. [3]). This graph is not meant for statistical purposes but rather to visually and descriptively capture marked theme instantiation. Users can hover over the colored sections of the graphs to show the marked theme type and how it was textually realized in their composition. On the whole, the graph summarizes where certain marked themes were used within the text and offers general insights into their frequency distribution. For example, a greater frequency of projecting themes towards the end of a scientific paper could reflect greater use of objective language in the results and discussion section; otherwise, the lack of modal themes in scientific texts could indicate the absence of subjective language through marked themes. Ultimately, this marked theme visualization should provide an initial impression of marked theme use, e.g., the use of the structural theme *in addition* or modal theme *obviously* and its realization in the text. This visualization is expounded upon further in its typification and semantic classification in the second results tab, as outlined next in Chapter 4.11.3.

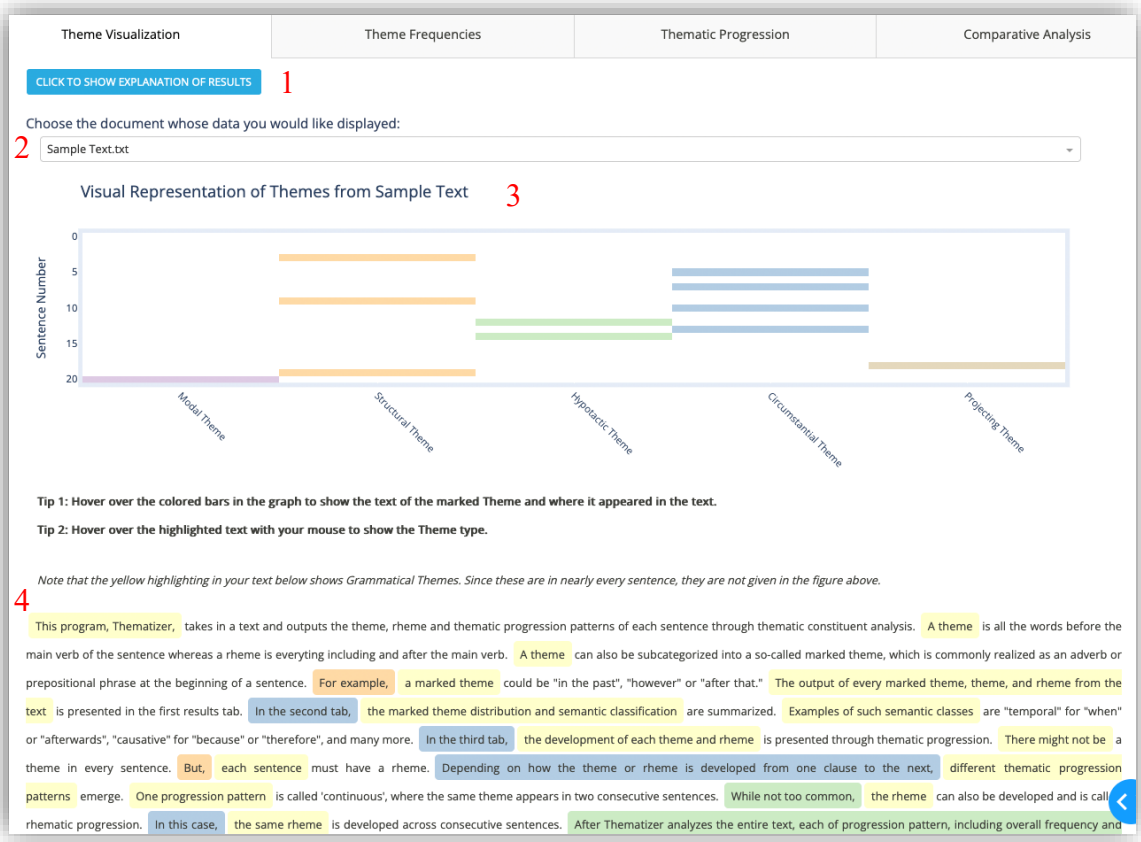


Figure 4-31: Sample output of Thematizer's first task and results in the first results tab Theme Visualization. *The diagram at the top summarizes where marked themes were realized within the text, whereas the highlighted text at the bottom is a visualization of all theme types realized in the user's text.*

The remainder of the first tab is comprised of the original text that the user input (cf. [4]). The output is complemented with highlights that mark the marked themes and grammatical themes, whose color refers to the theme type. For instance, hypotactic themes are highlighted in green, circumstantial themes in blue and grammatical themes in yellow. Rhemes remain unhighlighted. Hovering over the highlighted text reveals the thematic type so that the user does not need to refer to legend only. This is particularly useful for longer texts, which cause the user to scroll down beyond the legend.

Ultimately, the purpose of this tab is purely descriptive: the highlights and graph help to make the user's sentence structure and use of marked themes explicit. They bring to the forefront recurring structural tendencies that the user may have remained unaware of without such overt visualization and demarcation. Finally, these results form the analytical basis of the output presented in the remaining three tabs.

4.11.3 Results Tab 2: Marked Themes

In the second results tab, the marked themes' classification, frequencies and use in the input text are summarized. As shown in Figure 4-32, users can select the text from the first drop-down menu to summarize its marked theme analyses. If multiple documents have been uploaded, users can select multiple documents simultaneously for comparative purposes. The uploaded texts automatically populate the drop-down menu, even if only one text has been input.

In the second drop-down menu, the five marked theme types are listed: modal, structural, hypotactic, circumstantial and projecting themes. Once the user has selected a marked theme

type present in the text, the frequency of its semantic class is dynamically generated and output in the corresponding bar chart. If a text has no marked theme type that the user has selected, an empty graph appears with an indication of the marked theme's absence in red. The graph allows any number of texts and will dynamically resize depending on the number of texts considered; however, an upper limit of ten is recommended since the results within the graph become less legible with more than ten texts.

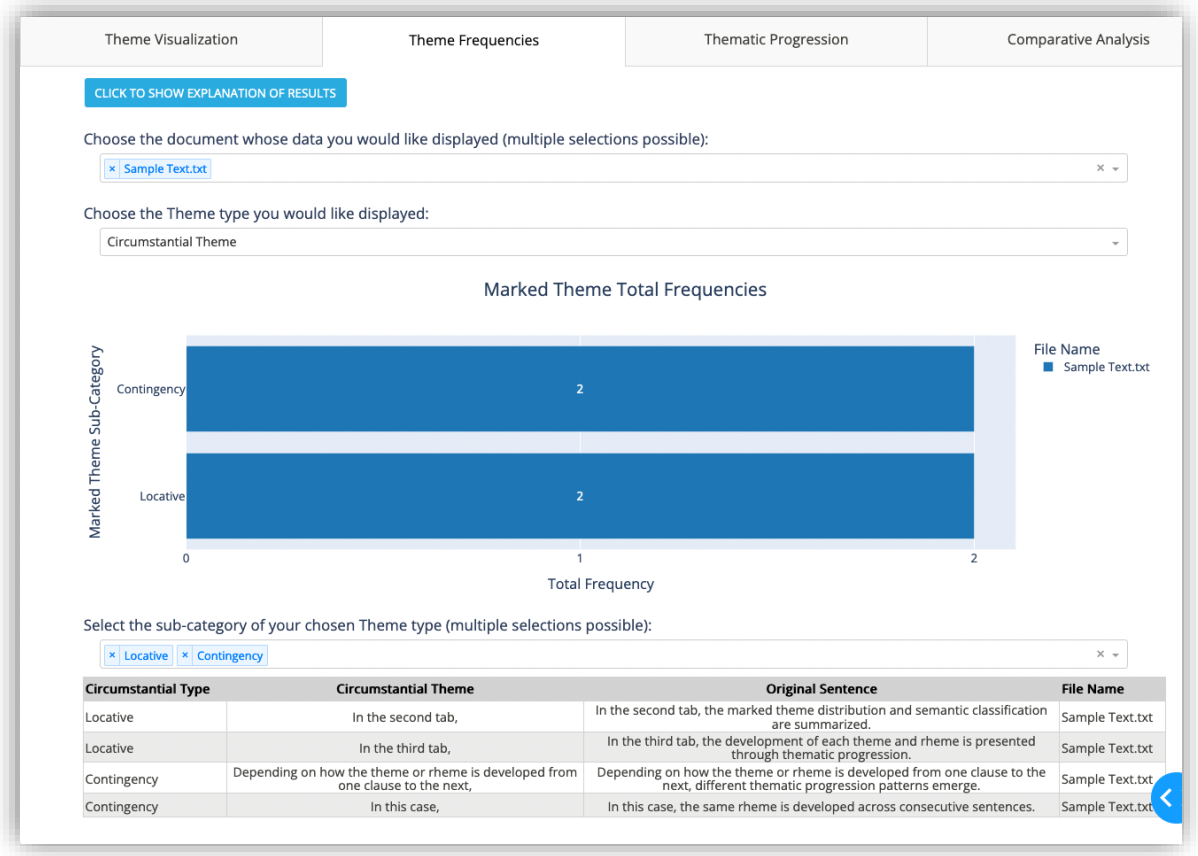


Figure 4-32: The second results tab in Thematizer summarizes the frequency of each marked theme type in the diagram at the top including the semantic subclassification of the marked themes present in the analyzed text. The table at the bottom extracts the examples of the marked themes as they were used in the original text. Here, multiple texts can be selected for comparative analyses.

Beneath the bar chart, Thematizer auto-populates the final drop-down menu based on which marked theme type the user selects. Since semantic classes are specific to the marked theme, the user may only select one marked theme type. However, multiple semantic classes may be chosen per marked theme type. This final drop-down menu then dynamically generates the table that appears at the bottom of the tab. Here, the marked theme that the user selected, its semantic class, its textual realization and the sentence in which it appeared are summarized. The file name where the text and marked themes were used is also appended to the table for the user's reference.

The purpose of this table is to automatically extract and summarize all instances of the selected marked theme in the user's text. This provides insight into which kind of marked themes the user employed, how they used them in their own text and at which frequency. This tabular summary can thereby highlight the variety of marked theme use. It can illustrate whether the user tends to employ the same marked theme or whether their text enjoys greater marked theme diversity. Additionally, the semantic class frequencies can shed light on the logical, rhetorical or text linguistic characteristics of the text, particularly from a comparative perspective.

4.11.4 Results Tab 3: Thematic Progression Analyses

The third results tab summarizes the results from the thematic progression analysis of the user's text (cf. Figure 4-33). As with the previous tabs, the user can select a specific text from a drop-down menu to show its analytical results. The user may select multiple documents to compare the frequencies of their thematic progression patterns and means of progression; otherwise, the results for a single text are presented as the default.

Figure 4-33: The text's thematic progression is summarized with the help of two diagrams and the input text in Thematizer's third results tab. The two diagrams show the frequencies of the thematic progression patterns and the means of progression. Beneath the graphs, the user is presented with their highlighted text, which shows how each sentence is thematically developed and by which means when hovering over the highlight with the mouse.

The first graph that appears beneath the drop-down menu is dynamically populated in terms of the text's thematic pattern frequencies, e.g., constant continuous progression, macrotheme or gapped continuous. The second bar chart summarizes the frequency of the text's means of progression, which indicates how thematic progression was instantiated. This can be coreference resolution, lexical repetition, discourse topic instantiation, paraphrase, existentials, clefts or thematic breaks.

Afterwards, the text that the user selected from the drop-down is reproduced at the bottom of the tab. Even if the user selects multiple texts, only the first text's output is printed. Multiple document selection therefore only applies to the frequency outputs for each bar chart. This is done to limit the amount of text presented on a single page, and is noted in the explanation of the results section of the tab. The highlights in the text reproduced here span the entire theme of each sentence, which includes any marked themes and the grammatical themes. The color of the highlight then indicates the thematic progression pattern across sentence clusters.

Users can receive more detailed information about the thematic progression pattern by hovering over the highlighted text. In the hover text, the progression pattern and means of progression are repeated. Also, the connecting elements that initiate the resulting thematic progression are given in the hover text. For example, if the lexeme *boxes* is repeated across two concomitant themes, then this token becomes the connecting element that instantiates constant continuous progression via lexical repetition.

This breakdown allows users to view the thematic progression results both with and within the text. Instead of having to rely on the statistical information from the two graphs and the legend that indicates thematic progression, the hover text complements and contextualizes the results where they are relevant within the text. Such visualization of the results overall and specific to their location within the text was motivated by deficiencies in previous research on the automated analysis of themes and thematic progression. The present approach aids in making thematic progression and its analysis within a text more comprehensible and tangible for users. Furthermore, the results in this tab can highlight how the user structures and develops their written discourse from sentence to sentence. Finally, the results here form the basis for the comparative, intertextual results that are presented in the final results tab.

4.11.5 Results Tab 4: Comparative Analyses

The fourth and final tab contains three bar charts that summarize the average frequencies of marked themes, thematic progression patterns and means of thematic progression for five text types: Wikipedia articles, L1 and L2 university texts, blog articles and lyrics. These averages are used as a comparison against the frequency values from the user's text. Just as in the previous tabs, the user can first select the input text from a drop-down menu if multiple texts are uploaded. If only one text is input for analysis, it is automatically selected for comparison. Then, the user can select any or all text types from the five provided to dynamically produce the average frequency bar charts below. Such a comparison allows for both intra- and intertextual analyses, i.e., analyses within single documents and text types and those across text types, as shown in Figure 4-34.

The purpose behind these three graphs is two-fold: First of all, users can readily compare the overall frequency of their own text's marked themes, thematic progression patterns and means of thematic progression to the average of texts with the same or similar text type. As intertextuality is a key indicator for a text's adherence to textual, structural, rhetorical or creative conventions, these graphs can provide evidence of whether their text deviates from or

falls within conventional use. Should the author wish to have their text be more reflective of such standards, users can make adjustments to their text to achieve values closer to the overall average. Conversely, if the writer wishes to deliberately deviate from these averages as a way to create a more idiosyncratic text, then the values from these graphs can also guide the writer in that direction.

Figure 4-34: The fourth and final results tab contains three diagrams whose data is a comparison of the user's text against five other text types (Wikipedia articles, L1 university essays, L2 university essays, song lyrics and blog articles). Specifically, the frequency of marked themes, thematic progression patterns and means of progression are summarized and juxtaposed in this tab's diagrams.

Secondly, the values from these graphs can help linguists confirm, refute or at least put into question findings from previous research on thematic progression patterns in various text types. While the average frequencies should not be the sole consideration in adherence to or deviation from texture characteristics of a text type, users can understand them as a subset of factors used for text typification. As the visualized data can be saved and incorporated into researchers' own investigations into text type and thematic inquiries, these analytical results can expedite the requisite text analyses.

4.11.6 Summary of Web Interface

Thematizer's web interface represents the program's frontend that enables both the functionality of and interaction with the text analyses and results. With the web interface, users can input their text for immediate and automated analysis. The results of this analysis can be saved for personal use and are visualized in tabs through highlighted text and summarizing charts.

Each of the four results tabs addresses a specific aspect of thematic analysis with the use of the user's text input. The themes and rhemes from the text input are delineated in terms of constituency and realization within the first tab. Marked themes, their frequency and semantic class are dynamically generated and summarized based on user input in the second tab. The third tab presents the results of the thematic progression analyses with the help of frequency bar charts and highlights that illustrate each sentence's progression pattern. The fourth and final tab tallies the frequencies of the thematic progression patterns, marked themes and means of progression from the user's text. These tallies are compared to average frequencies from five other text types for intertextual, comparative analysis.

The web interface, in summary, fulfills two functions: as a functional interface between the analytical backend and visualizing frontend on the one hand; as an interactable interface between the user and analytical output on the other. It is through this interdependent functionality that the underlying thematic development of the user's text is brought to the fore.

4.12 Summary of Methodology

The content presented in Chapter 4 reflects the theoretical framework that Thematizer rests upon and the methodological approach employed to implement the software as an automatic thematic parser. The two primary research questions of this work were first outlined as the impetus behind the research presented in this dissertation: deficiencies to overcome in previous thematic theory and computational models on the one hand; and facilitating accessibility to thematic structure through its operationalization by computational means on the other. These then informed the final thematic models for marked themes, the grammatical theme, rhemes and thematic progression patterns, which established the programmatic foundation of Thematizer.

The chapter then continued with the requirements, steps and materials for the developing, training and testing of Thematizer. Here, the programming libraries, application programming interfaces, text types and functional requirements for Thematizer were outlined in terms of their pertinence and functionality for automating thematic analysis.

The remainder of the chapter was dedicated to the core functionality of Thematizer in terms of its individual parsing steps, starting with text cleaning during pre-processing and ending with the presentation of the analytical results in Thematizer's web interface. The three core parsing steps that Thematizer performs for each text analysis were defined as the identification of theme and rheme spans, marked theme identification and classification, and thematic progression pattern classification. Each parsing task was detailed with respect to its purpose, its methodological approach from a thematic and computational perspective, and the final output that Thematizer yields. Thematizer's web interface concluded Chapter 4 as an illustration of its design, presentation and the sample output it produces upon thematic analysis of a text.

Chapter 5 – Results from Thematizer's Parsing Functionality

The present chapter outlines the performative results of Thematizer in terms of parsing accuracy and the prominent error classes to have occurred across its three parsing tasks. These results aim to highlight where Thematizer succeeded in its parses, where it failed and the error classes that pervaded each parsing task. The error classes specifically will provide insight into the resulting accuracy ratings that Thematizer achieved in its parses. In doing so, a fine-grained impression of Thematizer's performance and parsing capabilities throughout its three primary parsing tasks will come to light. An interpretation of how and why such error classes affected overall accuracy is reserved for the discussion in Chapter 6.

A presentation of Thematizer's performative results in the following will ultimately serve to provide initial answers to the present work's second research question: how can the operationalization of thematic structure via computational means make thematic structure accessible to writers? The answer to this question is then found in the degree to which Thematizer successfully operationalized thematic structure and provided accurate analytical output. As such, the quantitative results from Thematizer's parses can be used as a basis for answering the research question.

Chapter 5 is broken down into five sections. Chapter 5.1 first outlines how accuracy is defined and calculated in the present work. Then, Thematizer's overall accuracy with respect to the three parsing tasks – index identification, marked theme classification and thematic progression classification – is presented. The core error classes that affected Thematizer's accuracy are introduced afterwards to provide a general overview of the parsing difficulties that emerged while parsing. Chapters 5.2 to 5.4 examine each individual parsing task by outlining its accuracy and the core error classes that impacted Thematizer's parses. Each core error class is then reviewed in terms of its effect on parsing accuracy and the degree of pervasiveness in its respective thematic parsing task. Chapter 5.5 concludes the chapter with initial answers to the second research questions of this work. As finalized answers to the research questions are formulated in Chapter 6, the answers presented here form the foundation upon which subsequent conclusions can be drawn as key takeaways from the present research.

5.1 Thematizer's Overall Accuracy across All Thematic Parses

In this work, accuracy is defined by the rate at which Thematizer correctly performed each of the three core thematic parsing tasks. The important note here is that the accuracy rates in the individual steps are often cumulative. If the accuracy of a preceding step was compromised, then the accuracy of the subsequent step will likely have been as well. As such, errors that arise in earlier sections of the parse can become evident in latter portions.

In order to calculate Thematizer's accuracy, the present work employs the F-score, which is standard in classification tasks (Derczynski 2016). While Thematizer itself does not use classification techniques in a machine-learning sense, Spacy does rely on trained language models. These, in turn, affect the outcome of Thematizer's parses.

In determining the F-score for each parsing task, the parameters of precision and recall are used. Precision reflects the degree to which the program delivered correct classifications. This is expressed as the ratio of correct (true positive, T_p) classifications to both correct and incorrect (false positive, F_p) classifications, i.e.,

$$P = \frac{T_p}{(T_p + F_p)}$$

Recall, conversely, encapsulates the program's ability to identify and extract the relevant patterns in the parsing tasks. It considers the number of correctly identified classifications in relation to the number of unidentified or overlooked classifications. Recall is thus expressed as the ratio of correct (true positive, T_p) classifications to both correct and unidentified (false negative, F_n) classifications, i.e.,

$$R = \frac{T_p}{(T_p + F_n)}$$

Once both values have been calculated individually, they are combined into the F_1 score, which is the harmonic mean of precision *P* and recall *R*:

$$F_1 = \frac{2P \times R}{(P + R)}$$

This score thereby reflects both how well the program was able to correctly perform the parsing task (precision) and identify all relevant patterns present in the text (recall). In the present work, the F_1 score then refers to the overall accuracy of each parsing task that Thematizer performs. While convenient, this score can obscure actual precision and recall values since they are expressed in terms of a harmonic mean of precision and recall. Therefore, in the following discussions, the recall and precision values are also presented where appropriate to highlight the parser's classification accuracy in greater detail.

In order to substantiate the accuracy rates achieved here, accuracy rates from previous research are used for a baseline comparison. However, no previous work exists that performs identical thematic analyses as Thematizer does, which complicates comparisons of Thematizer's accuracy to previous models. Since the parses performed are a combination of tagging, syntactic and lexical entailment tests, composite accuracy rates from previous state-of-the-art research are used.

The first set of accuracy rates used is that of Spacy for its part-of-speech and dependency tagging as well as its pattern matching functionalities. These are reported to be 92.0% and 97.4%, respectively (Honnibal et al. 2020d). Next, the accuracy achieved in the thematic parsers functionally similar to Thematizer in the work by Domínguez et al. (2020) and Park & Lu (2015) was 74.0% and 93.0%, respectively. The average of these four sets of accuracy scores, 89.1%, was then used as a gold standard for the index identification and marked theme classification tasks. The use of this average accuracy for index identification and marked theme classification specifically is due to the tagging and syntactic nature of their parses.

For thematic progression tests, a combination of tagging, dependency and lexical entailment tests were required. For the latter, Roller et al. (2018) showed that cosine similarity tests for lexical entailment were able to achieve an average rate of 69.3% (2018: 5). This value was then averaged with the aforementioned accuracy rate for tagging and syntactic tests (89.1%) to

achieve a gold standard of 79.2%. It is argued that these composite accuracy rates per parsing task reflect the parsing complexity of each while being held to state-of-the-art standards that previous research has produced.

With a definition of accuracy and the baseline accuracy values established, Thematizer's overall accuracy can now be presented. The resulting F_1 scores for Thematizer's three parsing tasks with both training and test (i.e., validation) data are summarized in Table 5-1.

Parsing Task	Training Data F_1 Score	Test Data F_1 Score	Gold Standard F_1 Score
Theme/Rheme Index Identification	**85.8%**	92.0%	89.1%
Marked Theme Classification	94.9%	93.4%	89.1%
Thematic Progression Classification	80.2%	**75.9%**	79.2%
Thematizer's Final F_1 Score	85.7%	85.4%	79.2%

*Table 5-1: Individual and overall F_1 scores for each of Thematizer's three parsing tasks compared to F_1 gold standard scores from previous research. The final F_1 scores represent the composite accuracy rates comprehensively achieved in all three parsing tasks. Thematizer's F_1 scores that were below the gold standard are in **bold**.*

Considering Thematizer's final F_1 score first, which represents the combined accuracy rates of the three parsing tasks, Thematizer's accuracy reduced by 0.3% between the training and test datasets, yielding an F_1 score of 85.7% and 85.4%, respectively. The decrease in the **test** dataset was the result of the 75.9% F_1 score achieved in thematic progression classification despite the ≥ 90.0% accuracy in index identification and marked theme classification. While Thematizer was able to identify and classify marked themes with the training data at an accuracy rate similar to the test data, it was ultimately the index identification and thematic progression classification tasks that diminished the F_1 score for **training** texts (85.8% and 80.2%, respectively). Overall, however, Thematizer's final F_1 scores represent a composite increase in accuracy over the gold standard of 79.2% by at least 6.2% in both datasets.

Considering the individual accuracy rates for each parsing task, only two failed to reach the gold standard from previous research (cf. bold F_1 scores in Table 5-1). Compared to the 89.1% gold standard for syntactic and tagging tests, Thematizer achieved an F_1 score of 85.8% in the index identification parse for the training data alone. In thematic progression classification, Thematizer's test data parses yielded an accuracy of 75.9% compared to the gold standard of 79.2%. Only marked theme classification parses from both datasets were able to exceed the gold standard F_1 score.

Again, the difference in gold standard F_1 scores is derived from the kind of tests performed in each parsing task: index identification and marked theme classification required dependency and tagging parsing alone, which resulted in the average F_1 score of 89.1% in previous research. Conversely, thematic progression classification required dependency, tagging and lexical entailment parsing, which amounted to a gold standard of 79.2%. Thus, it was lexical entailment tests that decreased the gold standard F_1 score, a trend that is also visible in Thematizer's parses: parses requiring dependency and tagging tests alone achieved higher scores than those additionally requiring lexical entailment tests.

Depending on the text type, Thematizer was able to perform the three parsing tasks at varying levels of accuracy. To illustrate this range of accuracy, the heatmap in Figure 5-1 was created. There, the results for the F_1 scores from the individual text types from both datasets are presented according to their parsing task. Cells shaded in green indicate F_1 scores beyond the gold standard; those in yellow, orange or red represent F_1 scores below the gold standard.

F_1 Scores per Parsing Task w.r.t. Text Type				
Dataset	Text Type	Task 1: Index Identification	Task 2: Marked Theme Classification	Task 3: Thematic Progression
Training Texts	Wikipedia Articles	88.4%	93.2%	72.1%
	L1 University Texts	85.0%	95.2%	78.4%
	Blog Articles	86.8%	95.4%	84.0%
	Lyrics	85.9%	94.2%	88.1%
	L2 University Texts	78.0%	93.5%	84.9%
Test Texts	Gaming News Site	94.3%	97.8%	87.3%
	Newspaper Article	94.3%	94.1%	76.3%
	Linguistics Textbook	86.0%	92.6%	76.6%
	Reddit Comments	85.0%	89.2%	85.4%
	Editorial	91.8%	93.5%	73.1%
	Obituary	88.6%	94.1%	62.8%
	Blog Comments Section	94.6%	91.7%	84.7%
	Wikipedia Article	90.9%	95.1%	65.1%
	L1 University Text	91.0%	93.3%	75.9%
	Short Story	92.7%	95.7%	66.3%

Task 1 & 2 Legend	70.0%	80.0%	90.0%	100.0%

Task 3 Legend	60.0%	70.0%	80.0%	90.0%	100.0%

Figure 5-1: F_1 scores for Thematizer's three thematic parsing tasks with respect to text type. The legend for tasks one and two is based on the gold standard of 89.1% for tagging and dependency tests. The legend for task three is based on the comparative baseline accuracy of 79.2% for tagging, dependency and lexical entailment tests.

Compared to the scores from Table 5-1, there are considerably more F_1 scores below the gold standard for each text type. In the first parsing task, Thematizer yielded accuracy rates less than the gold standard of 89.1% in all training texts. This reinforces the particular difficulty Thematizer had with index identification in training, as the F_1 score from Table 5-1 first indicated (cf. $F_1 = 85.8\%$). When identifying theme/rheme indices in the test dataset, parsing the text from the linguistics textbook, Reddit comments and obituary revealed similar difficulties. That being said, Thematizer parsed the majority of test texts (seven of ten text types) with an accuracy of 90.9% or higher. This indicates an increase in parsing accuracy of at least 1.8% in the test texts compared to the gold standard (89.1%).

When classifying marked themes, Thematizer achieved an average of 94.3% and 93.7% accuracy for the training and test texts, respectively. While marginal, difficulties in parsing the test texts proved more common due to Thematizer overlooking marked themes present in the texts. Specifically, Thematizer's accuracy suffered most with the Reddit comments text from the validation dataset (cf. $F_1 = 89.2\%$), which was the lowest F_1 score of all text types in marked theme classification. This was on account of modal vocatives and interjections remaining unaccounted for. Ultimately, Thematizer proved to be most adept at parsing marked themes by exceeding the 89.1% gold standard in every text across both datasets.

For the third and final parsing task, Thematizer achieved the lowest F_1 scores. In fact, the thematic progression parse was the only task that Thematizer was unable to achieve an accuracy of at least $F_1 = 90.0\%$. For the training dataset, only blog articles, lyrics and L2 university texts exceeded the 79.2% gold standard. The test dataset, on the other hand, suffered considerably in terms of accuracy, failing to achieve the gold standard in seven of the ten test texts. These findings illuminate the nearly 5.0% difference in thematic progression classification accuracy

between both datasets as outlined in Table 5-1 (cf. 80.2% for training, 75.9% for test datasets). In fact, an F_1 score less than 70.0% was evident in the obituary, Wikipedia article and short story, which indicates the impact that lexical entailment had on parsing complexity.

In consideration of all three parsing tasks, Thematizer was able to successfully identify theme/rheme indices and classify marked themes 9.6% and 15.9% more accurately than thematic progression, respectively. As the former requires tagging and dependency tests for parsing, compared to the lexical entailment component for thematic progression classification, the results suggest Thematizer's strength in the syntactic, rather than the semantic, realm.[23] The diminished accuracy for thematic progression classification additionally suggests that cosine similarity and its testing parameters as a method for determining lexical entailment was insufficient.

To provide further evidence for these findings, the errors responsible for reducing Thematizer's parsing accuracy will be addressed. Altogether, 23 separate errors classes emerged across the datasets and were shared among one or more of the three parsing tasks (see Figure 5-2 for a summary of each error class; for a breakdown of each error class according to parsing task and dataset, see Figure A-1 and Figure A-2 in Appendix A). A close examination of the exact F_1 scores together with their accompanying error cases will elucidate their effect on Thematizer's overall accuracy in the respective parsing task. As the errors in Figure 5-2 are core, recurring errors throughout all parsing tasks, reference to these error cases will be made throughout the remainder of the present chapter and in finer detail throughout Chapter 6.

Cases that accounted for more than 10.0% of the parsing errors were lexical entailment (29.9%), coreference resolution (17.4%), t-unit parsing (10.7%) and subject/root index identification (10.2%). Lexical entailment, which covers hypernymy, hyponymy, meronymy, synonymy, antonymy, paraphrase and ellipsis, was unique to the thematic progression classification task and contributed most to parsing deficiencies in both training and test datasets. Lexical entailment is resolved through Thematizer's cosine similarity tests as one means of progression across sentences (cf. Chapter 4.10.5). As suggested above, it is this error case specifically (but not solely) that led to the lower F_1 scores in thematic progression classification for training (80.2%) and test texts (75.9%). Nearly a third of all error cases belonging to lexical entailment illustrate the questionability behind employing cosine similarity tests as one means to resolve thematic progression.

The second most common error class, coreference resolution, represents Thematizer's difficulty with tracing coreference chains between concomitant sentences on the basis of the coreference chains returned by Coreferee's parse (cf. Chapter 4.10.2). Such errors emerged during all parsing tasks but affected thematic progression classification most. There, Thematizer failed to account for coreference resolution as a means of thematic progression due to erroneous indexing and faulty resolution through personal pronouns. Conflating the dummy-*it* in clefts and projecting themes with coreferential proforms further accounted for coreference errors. For test texts specifically, coreference resolutions proved to be nearly as problematic as lexical entailment, as evidenced by the thicker arm in Figure 5-2.

[23] It should be reiterated that merely the causes of the misparses are presented in the current results sections alone. *How* these errors came about and *why* they happened will be thoroughly explained in Chapter 6. The presentation of the causes here serves as an initial explanation of their effect on the resulting accuracy and an introduction to the individual error cases outlined in Chapter 6.

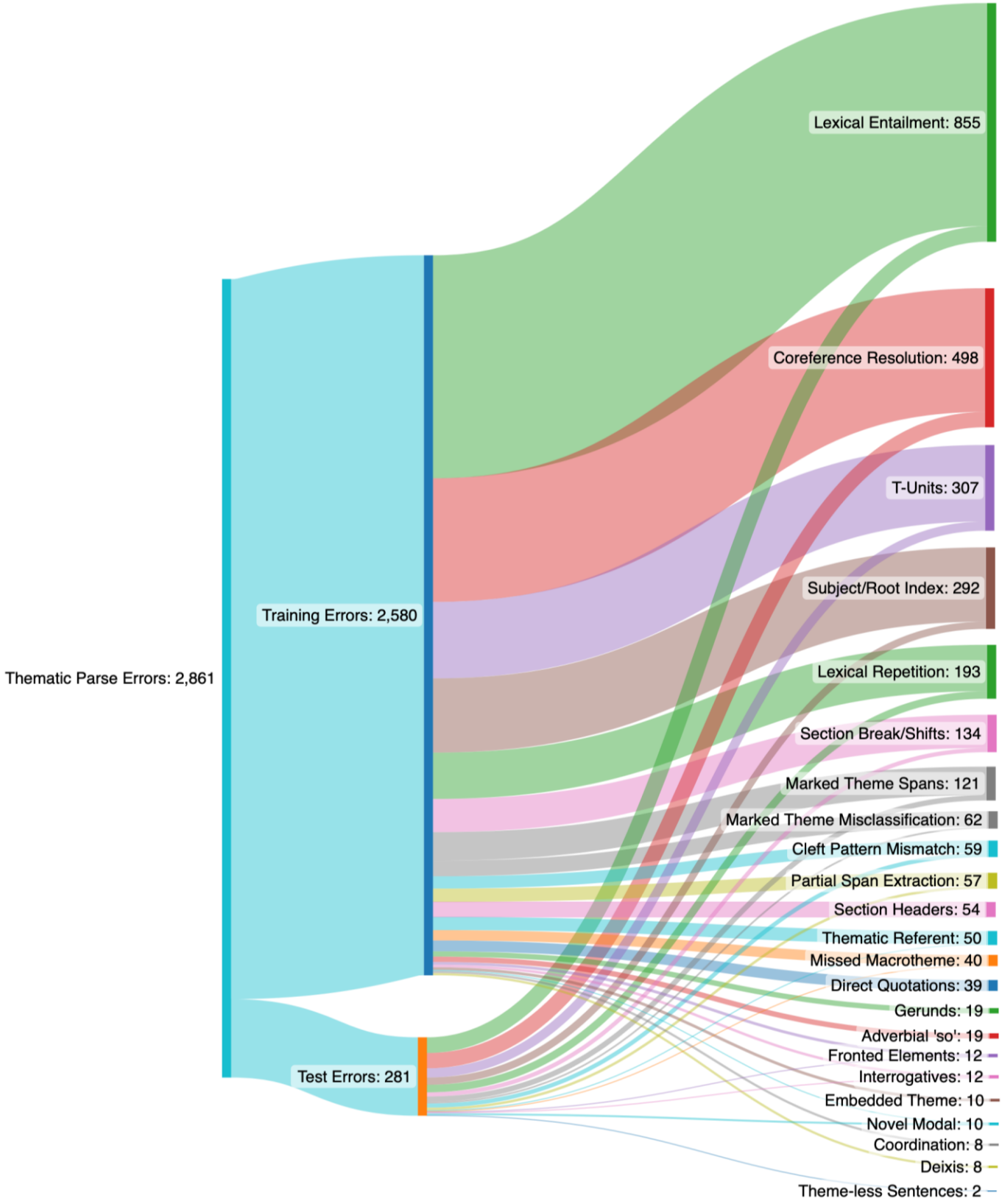


Figure 5-2: Summary of all error cases that emerged in training and test texts during thematic parsing. The width of the flows (arms) and the absolute frequencies indicate the degree to which the error permeated the parses. Shared error cases are indicated by the flows and connecting nodes between training and test errors.

Similar to coreference errors, misparsed t-units affected each of the three parsing tasks. Thematizer was programmed to split concomitant independent clauses separated by conjunctive adverbials, coordinating conjunctions, semicolons, colons and hyphens. If one of the concomitant clauses was dependent, then the two clauses should not have been split from one another. T-unit misparses were thus a result of either not splitting concomitant independent clauses or erroneously splitting dependent clauses from their independent clause. The frequency of t-unit misparses at 10.7% thereby indicates the difficulty Thematizer had in correctly parsing independent and dependent clauses, particularly since this error pervaded each parsing task.

Subject/root index identification refers to Thematizer's ability to identify the correct grammatical subject and congruent finite verb in the independent matrix clause. With these indices, the marked themes, grammatical theme and rheme spans were extracted and split from one another. Errors from this task were most readily due to comma misplacement, clause dependency misparses, disambiguation and coreference resolution. While Thematizer had less difficulty with index identification when parsing the test texts (as indicated by the thinner arm in Figure 5-2), considerably more misparses affected Thematizer's analysis of training texts. This tendency is reflected in the lower F_1 score achieved with training texts (85.8%) compared to the test texts (92.0%) in the index identification task.

Coreference, t-unit and subject/root index misparses were especially denigratory in that they had a cascading effect on subsequent parsing tasks and resulting accuracy. If these errors emerged during index identification, then Thematizer likely misparsed or misclassified marked themes in the second parsing task. In the third parsing task, the selfsame errors then caused thematic progression classification to either fail or only be partially correct due to mismatched thematic referents. Thus, it was not only the frequency of these misparses but also their compounding nature that affected the ultimate F_1 scores of each parsing task.

The remaining 19 error cases presented in Figure 5-2 occurred at frequencies less than 10.0% and ranged between a maximum occurrence of 6.7% and a minimum of 0.1%. Despite their lower individual frequencies, together they constituted 909 of the 2861 error cases across training and test datasets. Their contribution to a reduction in the parsing accuracy of each task is therefore of considerable yet variable consequence. Further, their emergence is an additional result of the cascading errors originating from t-unit and subject/root index misparses in Thematizer's first parsing task. Although the errors of section headers, direct quotations, gerunds, the adverbial *so*, coordination and embedded themes were unique to training texts, this does not unilaterally preclude their potential presence in novel texts. The errors shared between both datasets does, however, lend credence to the parsing difficulty Thematizer experienced with each of the error cases.

Across all text types, the pervasiveness of errors thereby reflects the difficulty Thematizer experienced in operationalizing particular thematic structures. If clefts were a common occurrence in the text and commonly misparsed, as was the case with the linguistic textbook, this shows how Thematizer failed to consistently trace and capture cleft structures in its parses. What this means is that the programmatic approach used to identify clefts as an operationalization of thematic structure was partially deficient. The greater the error frequency across parses, the more the operationalization of thematic structure should be scrutinized.

This finding can thus serve to initially answer the present work's research question on the accessibility to thematic theory via its operationalization by programmatic means. Thematizer's overall accuracy of 85.7% for training texts and 85.4% for test texts exceeds the 79.2% gold standard by at least 6.2%. Yet, considering the accuracy of the three parsing tasks across both

datasets individually, the index identification accuracy of 85.8% in training texts failed to reach the gold standard of 89.1%. Similarly, in the test dataset, the F_1 score for thematic progression classification was the lowest of all parsing tasks at 75.9%, which is 3.3% less than the gold standard. Marked theme classification alone was able to exceed gold standard accuracies in both datasets. Therefore, while operationalization of thematic theory was entirely successful in marked theme classification, only partial success can be claimed for index identification.

Where operationalization should be scrutinized most is in the parses from thematic progression classification, which achieved the lowest F_1 scores of all three parsing tasks (training texts: 80.2%; test texts: 75.9%). In fact, the reduction in parsing accuracy in the test dataset to less than the gold standard F_1 score lends greater credence to the questionability behind how thematic progression classified was approached in the present work. Furthermore, the high degree of errors in this parsing task sheds potential doubt on the reliability of Thematizer's thematic progression output. Therefore, operationalization of thematic theory for thematic progression remains deficient.

The frequency of parsing errors as a function of Thematizer's parsing accuracy is thus indicative of the degree to which thematic theory was successfully operationalized. For that reason, the next sections will take a closer look at the three primary parsing tasks and their error distribution. Doing so will shed light on which error cases emerged in which parsing task from both datasets. This will then provide further evidence and contextualization for the answers to the degree of operationalization behind thematic theory.

5.2 Error Cases and Accuracy Rates for Index Identification

The first parsing task that Thematizer performs is the identification of indices used to determine marked theme spans, grammatical theme spans and rheme spans. The grammatical subject of the matrix clause was used together with its congruent root index to determine the boundary between the grammatical theme and rheme. Should sentence constituents appear in front of the grammatical subject and be dependent on the verbal root, then they were denoted as marked themes. The index at the end of the marked theme phrase was then used to demarcate the boundary between the marked theme and grammatical theme.

If Thematizer either failed to extract the exact span or it missed the requisite spans, then the resulting parse was considered incorrect. The former error type falls under precision errors whereas the latter contributes to Thematizer's recall score. As Thematizer had to potentially identify three different spans – one for marked themes, one for the grammatical theme and one for the rheme – there was potential for multiple misparses in a single sentence.

How well Thematizer was able to identify the various spans is summarized in Figure 5-3, which includes the F_1 score broken down into its respective precision and recall value for each text type. Since tagging and syntactic parses alone are required for the index identification task, the gold standard of 89.1% from previous research was used for comparison against Thematizer's accuracy.

First of all, Figure 5-3 shows that recall suffered most with the training texts in this first parse. In other words, Thematizer more frequently overlooked the relevant theme and rheme spans in training texts. This was largely due to grammatical errors in the original that resulted in erroneous spans. In fact, none of the recall scores for training texts reached the 89.1% gold standard that previous parsing tools were able to achieve. On average, only 80.2% of all spans present in the training texts were extracted. Where extracted, 89.4% of the spans were then

demarcated with the correct indices on average. Therefore, so long as Thematizer was able to extract the relevant span within the training text, the resulting span indices were correctly nearly 90.0% of the time.

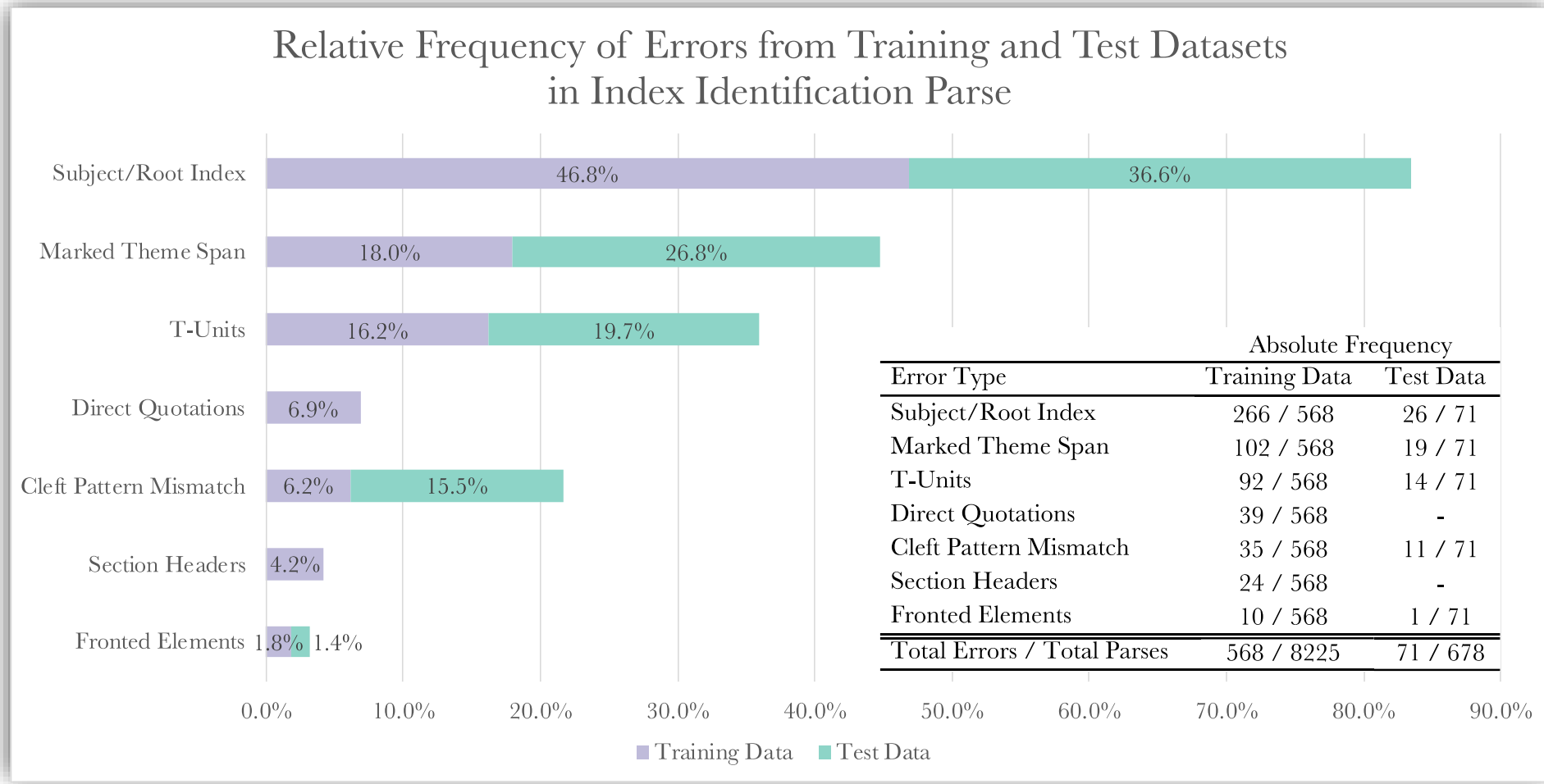
Index Identification Parsing Accuracy w.r.t. Text Type				
Dataset	Text Type	F_1 Score	Precision	Recall
Training Texts	Wikipedia Articles	88.4%	92.3%	84.9%
	L1 University Texts	85.0%	88.6%	81.6%
	Blog Articles	86.8%	87.3%	86.2%
	Lyrics	85.9%	90.3%	81.8%
	L2 University Texts	78.0%	88.6%	69.6%
Test Texts	Gaming News Article	94.3%	93.5%	95.1%
	Newspaper Article	94.3%	93.3%	95.4%
	Linguistics Textbook	86.0%	82.2%	90.2%
	Reddit Comments	85.0%	81.0%	89.5%
	Editorial	91.8%	91.8%	91.8%
	Obituary	88.6%	87.5%	89.7%
	Blog Comments Sectio	94.6%	93.8%	95.3%
	Wikipedia Article	90.9%	90.2%	91.7%
	L1 University Text	91.0%	91.6%	90.5%
	Short Story	92.7%	92.7%	91.8%
Legend	60.0%	70.0%	80.0%	90.0%	100.0%

Figure 5-3: Accuracy rating broken down into F_1 score, precision and recall for Thematizer's first parsing task, index identification, for extracting the marked theme, grammatical theme and rheme spans.

For the test texts, all recall scores surpassed the gold standard and achieved an average recall rate of 92.1%, an increase of 2.1% compared to the training texts. Precision, on the other hand, was less successful, such that only 89.8% of the extracted spans were demarcated correctly. Despite the lower precision score for test texts compared to recall, a minimal increase in average precision by 0.4% over the training texts was achieved. This was due to the majority of the test texts achieving precision scores greater than 90.0%, which counterbalanced the lowest precision values from the linguistics textbook and Reddit comments parses.

It was this set of test texts as well as the L2 university texts from the training set that experienced the greatest number of index identification errors. In the case of L2 university texts, incongruence between subjects and verbs, sentence fragments and incomplete sentences were the cause behind Thematizer's inability to identify spans and their indexical boundaries. Grammaticality, therefore, formed the greatest source of parsing errors in this specific training text type. For the text from the linguistics textbook, non-projecting clefts were commonly misparsed as coreferential structures, and ellipses caused Thematizer to incorrectly define the index of the grammatical subject of the matrix clause. Finally, numbered lists were a common occurrence in the Reddit comments text, which Thematizer frequently parsed erroneously as the grammatical subject and, by extension, grammatical theme.

These errors, while particularly prevalent, were not unique to these text types. Rather, these and additional errors affected each text type to varying degrees and can help to explain each of the precision and recalls scores. The error classes that emerged in both datasets during index identification are summarized in Figure 5-4, which are a subset of all error classes originally presented in Figure 5-2.



	Absolute Frequency	
Error Type	Training Data	Test Data
Subject/Root Index	266 / 568	26 / 71
Marked Theme Span	102 / 568	19 / 71
T-Units	92 / 568	14 / 71
Direct Quotations	39 / 568	-
Cleft Pattern Mismatch	35 / 568	11 / 71
Section Headers	24 / 568	-
Fronted Elements	10 / 568	1 / 71
Total Errors / Total Parses	568 / 8225	71 / 678

Figure 5-4: Relative and absolute frequencies of recurring and core error cases from Thematizer's first thematic parse, index identification. Data from both the training and test data set are presented for comparison.

The error frequencies presented in the figure are not a summary of the total number of correct parses versus incorrect parses. Instead, the percentages indicate the frequency of the given error class with respect to the other error classes that occurred. Absolute frequencies help to elucidate the exact frequency of the error cases belonging to this parsing task. Total errors provided with the absolute frequencies are precision errors (false positives) only; false negatives are not included in the tallies.

Of the 8225 parses performed for the training data, a total of 568 was erroneous; for the test data, a total of 71 from the 678 total parses was erroneous. In both datasets, the vast majority of the errors was the misidentification of the subject and finite verb root in the matrix clause (46.8% for training texts and 36.6% for test texts). Both of these indices were used to demarcate the grammatical theme span and the rheme span. Partial extraction of the grammatical theme or the inclusion of elements not belonging to the grammatical theme ultimately constituted subject/root index misparses. Compared to the test datasets, misparses of the subject/root indices from the training texts increased by 10.2%.

By and large, difficulties with determining the theme and rheme indices were largely due to clause dependency misparses and comma misplacement. The absence of commas after dependent, introductory clauses contributed most to a reduction in accuracy rates across all text types. For instance, in *When they were young, music was banned*, the comma separates the introductory clause *when they were young* from the independent matrix clause *music was banned*. For Spacy, such commas critically function as a boundary indicator and aid in determining syntactic dependencies of the elements within the introductory clause. Where superfluous or missing, Spacy was unable to accurately identify clause borders and dependency parses. Thematizer, in turn, failed to deliver correct theme and rheme spans, ultimately resulting in diminished accuracy rates.

Aside from grammatical theme spans and rheme spans, Thematizer was also tasked with demarcating the marked theme spans during the first thematic parse. Elements appearing in front of, but syntactically independent from, the grammatical subject were delineated as a marked theme for subsequent parsing separate from the grammatical theme. While marked theme classification achieved the highest accuracy of all three parsing tasks in both datasets, correct identification of their exact indexical spans was not without error. Test datasets, in particular, exceeded the training data's error frequency of 18.0% for marked themes and reached a total of 26.8%. A higher frequency of dependency misparses due to comma misplacement, fronted elements and lexical repetition was at the root of the test data errors. In turn, Thematizer overlooked or failed to delineate the entire marked theme span, which additionally accounts for the increase in error frequency of marked theme spans by 8.8% for the test datasets.

A similar trend can be seen in the misparse of t-units, which constituted 16.2% of the errors in training texts and 19.7% in test datasets. Two parsing errors for this error class are amalgamated in their relative frequencies: firstly, the incorrect splitting of dependent clauses from their independent matrix clauses; secondly, not splitting two concomitant independent clauses conjoined by conjunctive adverbials, coordinating conjunctions, semicolons, colons or hyphens. Ultimately, it was the latter case that contributed most to the t-unit errors, which compounded parsing errors when classifying marked themes and thematic progression.

The reason for these errors is due to how Thematizer identifies multiple verbal roots: if multiple finite verbs exist across dependent and/or independent clauses, only the last instance of the finite verb is considered the root. This causes Thematizer to erroneously include the entire

previous independent clause in the grammatical theme as marked in bold in (1). There, Thematizer marked all constituents up to the second finite verb as the grammatical theme, when the two independent clauses should have been split and received their own thematic constituent analyses.

(1) **Errors emerged in the analysis; therefore, the accuracy** suffered.

Direct quotations followed a similar programmatic approach to t-units, whereby they were separated from the independent projecting clause if the direct quotation itself was an independent clause (e.g., *I said, "I don't want to eat."* has two independent clauses to be separated at the comma). While direct quotations were present in both datasets, only the training dataset experienced misparses at a rate of 6.9%. Parsing direct quotations took place during pre-processing (cf. Chapter 4.7.1), whereby independent quotations were split from independent projecting clauses. If missed there, then Thematizer was unable to correctly identify the correct grammatical theme of both independent sentences. The error rate of 6.9% in training texts indicates that the programmatic approach to direct quotation parsing failed to account for all, potentially complex, expressions with quotes.

The error cases explained thus far – subject/root index, marked theme spans, t-units, direct quotations – represent a unique case, in that the misparse of one nearly invariably caused the misparse of another. If Thematizer failed to split t-units or direct quotations, then the subject and root indices were either missed entirely or misidentified. Where marked themes were also present in the sentence, identification of the incorrect subject and root indices caused Thematizer to demarcate the marked theme boundary incorrectly. These misparses were ultimately the result of incorrect dependency parses returned by Spacy on account of incorrect commas, clause dependencies, disambiguation and coreference resolution.

The interdependency amongst these parsing cases is thus another example of the cascading effect a misparse can have within a single parsing task. This group of errors is of particular importance because it not only reduced the precision and recall values for the index identification task alone; it also affected the accuracy rates of the subsequent parsing tasks since they are dependent on the index spans produced during index identification. Such errors can be considered a root-cause error, such that their emergence inevitably affected the remaining thematic parses.

Returning to the remaining error cases from Figure 5-4 above, cleft pattern mismatches affected both datasets at similar frequencies, although test texts experienced a greater number of cases (cf. 6.2% for training texts and 15.5% for test texts). Where errors occurred with clefts was through a mismatch in the pre-defined cleft pattern that Thematizer used when perusing the text through token realization and dependency parses. Commonly, adverbials were inserted within cleft patterns, e.g. *It has, however, been shown that [...]*. The inserted adverbial deviated from the predefined dummy-*it* + *copula* + *participle* + *that*-clause, which caused Thematizer to ignore the cleft. Furthermore, due to the inserted conjunctive adverbial, Thematizer assumed independence in the projecting *it is* and the projected *that*-clause, which contained its own subordinate subject and congruent verb. Finally, the pre-defined pattern for non-projecting clefts was often conflated with coreferential *it is* structures. Therefore, Thematizer erroneously identified the coreferential *it* as a dummy-*it* and cleft with an adjectival or infinitival predicate. As clefts were a common occurrence, particularly in more formal, academic registers, such text types exhibited more frequent cases of mismatched and overlooked cleft structures.

The final two error cases from index identification were misparsed section headers and fronted elements. Since the test texts had no section headers, this did not pose a problem during parsing. However, the 24 cases of section header misparses in training texts indicate that this problem may still appear in novel texts that Thematizer was not trained on. Most commonly, Thematizer failed to identify section headers as dependent clauses both structurally and functionally separate from text segments that followed. Although section headers commonly had no finite verb, Thematizer assumed the noun phrase to be the grammatical subject and thereby the grammatical theme. Span identification was then based on this false assumption, which then led to fragments, i.e., text segments without a finite verb and/or grammatical subject, being thematically parsed. Instead, these should have been parsed as a rheme only to initiate thematic breaks or rhematic progression during later parsing.

Fronted elements were the least common error class to occur in both datasets and comprised vocatives and colloquial abbreviations, such as *ty* for *thank you*. Such misparses amounted to a total of ten cases for the training texts and a single case for the test data. While uncommon, the misidentification of fronted elements had a similar effect that misparsed t-units and subject/root indices had: their error during the first parsing task propagated to subsequent thematic parsing tasks. As such, their emergence, while minimal, affected the resulting parsing accuracy of all three parsing tasks.

In light of these results, the preponderance of index identification misparses stemming from subject/root dependency misparses indicates the added complexity in requiring a grammatical theme to be separated from marked theme and rheme spans where present. This caused Thematizer's parsing accuracy with the training texts to reach 85.8%, which is less than the gold standard of 89.1%. While Thematizer yielded an F_1 score of 92.0% for the test dataset, the sample size was a total of ten texts only, which limits the representativity and generalizability of the F_1 score. For that reason, it cannot be conclusively stated that thematic structure in terms of index identification was successfully operationalized. In order for that to be the case, Thematizer would have had to achieve greater parsing accuracy with the training texts and/or achieve a similarly high F_1 score in a sample size of texts of at least 30 texts for generalizability. As syntactic, dependency tests alone were used as the parsing methodology for index identification and yielded the greatest number of errors, this programmatic approach may be at the root of Thematizer's lower accuracy for its first parsing task.

5.3 Error Cases and Accuracy Rates for Marked Theme Classification

Following the same structure as Chapter 5.2, the present section considers the parsing results of the marked theme classification task in terms of its accuracy. The F_1 scores for each marked theme from both training and test datasets are first introduced to provide a general impression of how well Thematizer classified the five marked theme types. Afterwards, the most common error classes that emerged in each marked theme parse are summarized as a reflection of their effect on parsing accuracy. A brief listing of the causes of these error classes together with the accuracy results concludes the present section and highlights the key findings from the marked theme parses.

The accuracy of marked theme classification was affected and determined by three factors: First, whether the marked theme spans were correctly extracted in their entirety; second, whether the marked themes were correctly classified into one of the five marked theme classes; and third, whether the marked theme was correctly classified into its corresponding semantic subclass. As touched upon in the previous section, if Thematizer incorrectly identified a marked theme or missed the marked theme entirely in the index identification task, the misparse would carry

over to this task as well. Such errors were also considered in the determination of Thematizer's parsing accuracy of marked themes.

With these conditions in mind, the F_1 scores that Thematizer achieved in parsing each marked theme type are summarized in Figure 5-5. Due to the fewer number of errors in marked theme classification, F_1 scores alone are presented in the following discussion. The precision and recall scores for each marked theme class can be found in Table A-1 in Appendix A.

With the exception of modal themes, Thematizer was able to achieve F_1 scores of 90.0% or greater for all marked themes in both training and test texts. Differences in parsing accuracy between both datasets were approximately 2.0%, with modal themes alone possessing the greatest difference in parsing accuracy of 20.7%. While modal themes were a rare occurrence in both data sets (163 individual cases from the 3565 total marked themes in the training texts compared to 21 cases from the 301 total marked themes in the test texts), novel modals that Thematizer had not been trained on caused half of the modal themes to be missed in test texts. This ultimately caused the F_1 score for test texts to suffer.

Another trend evident from Figure 5-5 is a decrease in parsing accuracy between training and test datasets for structural and projecting themes. Here, failure to split t-units into two separate independent clauses joined by coordinating conjunctions in test texts resulted in greater parsing errors. For projecting themes in training texts, both a failure to identify projecting themes and the assumption of non-projecting clefts instead of coreference led to a reduction in parsing accuracy.

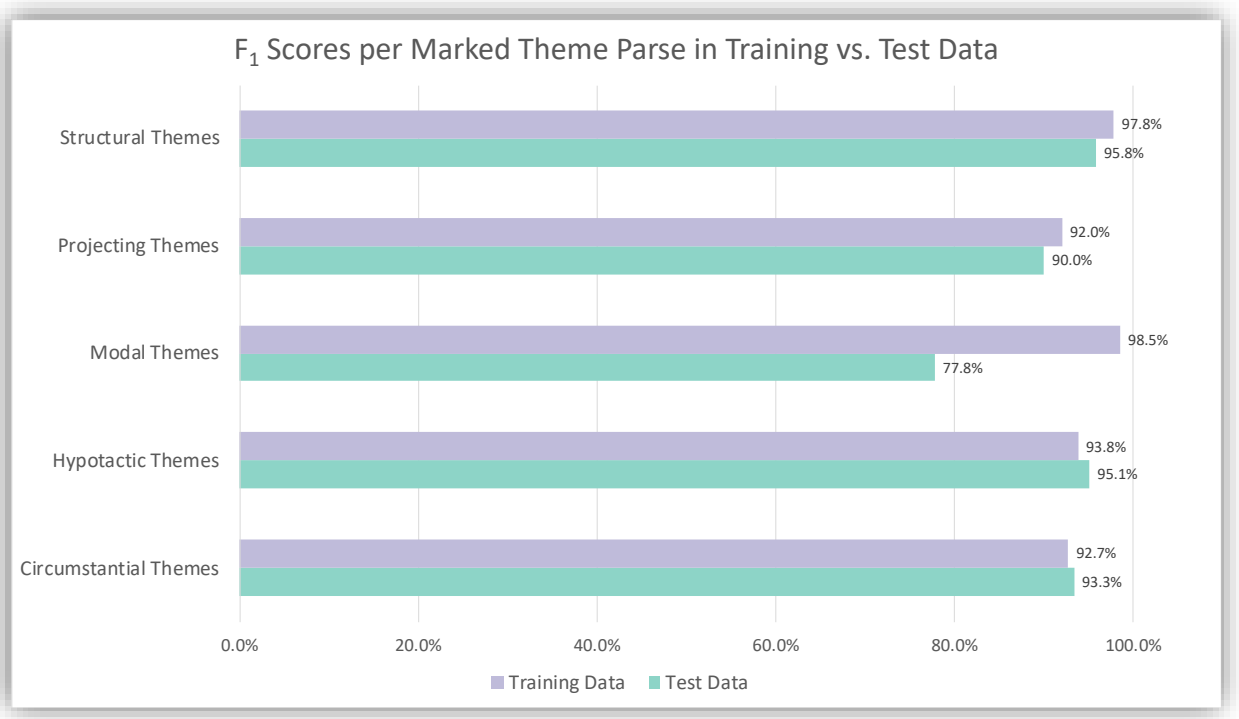


Figure 5-5: F_1 scores for each of the five marked theme types that Thematizer is tasked with parsing and classifying during its second thematic parsing task. Results from both training and test datasets are provided for comparison.

At the same time, an improvement in parsing accuracy between datasets is reflected in the F_1 scores of hypotactic and circumstantial themes. Despite their parsing complexity, an increase in their parsing accuracy can be viewed as an even more successful operationalization of

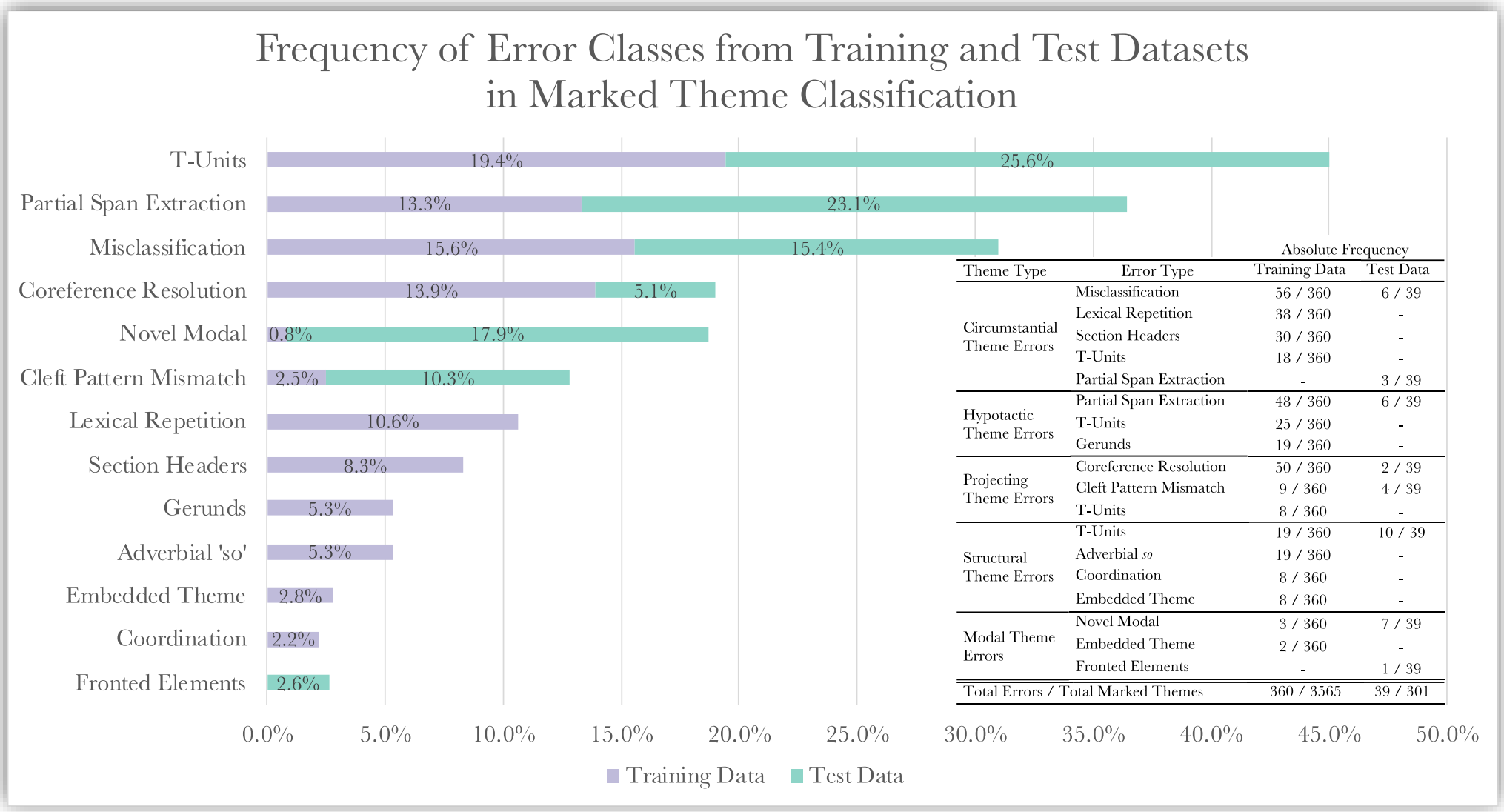
hypotactic and circumstantial themes through Thematizer. In fact, the high F_1 scores for all marked themes but modal themes provide evidence for the programmatic approach and its underlying theoretical model employed for parsing marked themes. As the F_1 scores exceeded the gold standard of 89.1%, this, too, indicates an improvement to thematic parsing over previous automated parsing tools.

In total, marked theme classification experienced the fewest parsing errors across the three thematic parsing tasks, totaling 360 for the training texts and 39 for the test texts. The exact error classes and their frequencies are summarized in Figure 5-6 with both relative and absolute frequencies provided for a more concrete comparison between both datasets. Further, a breakdown of the absolute error frequencies per marked theme type illustrates where the specific errors emerged and at what frequency.

Errors stemming from t-unit misparses constituted the greatest number of errors across both datasets. These were a byproduct of t-unit misparses from the index identification parse and resulted in sentence constituents being parsed as a marked theme despite not syntactically belonging to a marked theme class. Modal themes were the only marked theme class whose parse was not affected by t-unit errors. Otherwise, hypotactic themes were affected most by t-unit misparses, followed by structural themes, circumstantial themes and finally projecting themes (errors totaling 25, 19, 18 and 8, respectively). Additionally, the greater number of t-unit misparses from the index identification task in the test texts is reflected in the correspondingly high frequency of t-unit errors for test texts here during marked theme classification. This finding reinforces the cascading effect that misparses during the first thematic parsing task have on subsequent parsing tasks.

Partial span extraction, whereby sentence constituents from the marked theme span were left out, was the second most common misparse and occurred 9.8% more frequently in test than in training datasets (13.3% for training texts versus 23.1% for test texts). This error affected hypotactic themes in both datasets as well as few circumstantial themes in test texts. The cause of these misparses can be traced back to so-called right edge dependents and their syntactic heads. In hypotaxis in particular, these were required to determine the end of the marked theme span, e.g., *out* in *because it was still dark out*. Thematizer's false identification of the right edge in hypotactic themes specifically prevented some elements from being parsed as part of the marked theme.

Partial extraction in circumstantial themes occurred with temporal noun phrases that are dependent on subsequent prepositional phrases, e.g., *day* in *the day after the race*. Here, Thematizer extracted and classified the temporal noun phrase alone while ignoring the prepositional phrase due to erroneous right dependent parsing. As projecting themes do not require right edge dependents for parsing, this was not an issue for the marked theme class. Due to the right edge dependents always signifying the end of structural and modal themes regardless of syntactic complexity, Thematizer was consistently able to extract these marked themes in their entirety.



		Absolute Frequency	
Theme Type	Error Type	Training Data	Test Data
Circumstantial Theme Errors	Misclassification	56 / 360	6 / 39
	Lexical Repetition	38 / 360	-
	Section Headers	30 / 360	-
	T-Units	18 / 360	-
	Partial Span Extraction	-	3 / 39
Hypotactic Theme Errors	Partial Span Extraction	48 / 360	6 / 39
	T-Units	25 / 360	-
	Gerunds	19 / 360	-
Projecting Theme Errors	Coreference Resolution	50 / 360	2 / 39
	Cleft Pattern Mismatch	9 / 360	4 / 39
	T-Units	8 / 360	-
Structural Theme Errors	T-Units	19 / 360	10 / 39
	Adverbial *so*	19 / 360	-
	Coordination	8 / 360	-
	Embedded Theme	8 / 360	-
Modal Theme Errors	Novel Modal	3 / 360	7 / 39
	Embedded Theme	2 / 360	-
	Fronted Elements	-	1 / 39
Total Errors / Total Marked Themes		360 / 3565	39 / 301

Figure 5-6: Relative and absolute frequency of error classes from Thematizer's second parsing task, marked theme classification, across training and test datasets.

Continuing with the circumstantial theme errors, their misclassification proved to be the most frequent error class at approximately 15.0% for each dataset. Misclassification in this context refers to the semantic subclass that Thematizer categorizes the circumstantial theme into. What determines the semantic subclass is then the noun phrase that follows the preposition in the circumstantial. While class disambiguation was built into Thematizer for this specific task, errors invariably arose from noun phrase misparses as the realizational complexity of the circumstantial theme increased. This caused Thematizer to classify instances such as *on the morning of my wedding* as LOCATIVE despite the temporal noun *morning*. Since disambiguation was not required for the other marked theme types, they experienced no misclassifications.

Coreference resolution misparses frequently stemmed from cleft pattern mismatches and are reminiscent of the same error class from the index identification parse. During marked theme parsing, projecting themes were either conflated with coreferential structures or missed due to inserted conjunctive adverbials within the projecting theme. In the first case, Thematizer assumed coreference with the proform *it* and subsequent copula, which overwrote the pre-defined cleft pattern used in searching for clefts within the text. This, in turn, prevented Thematizer from identifying the construction as a marked projecting theme. In the second case, conjunctive adverbials such as *however* or *thus* were inserted into the projecting structure, which prevented a dependency and structural match with the pre-defined projecting theme pattern. For example, the projecting theme *it is, however, clear that…* failed to be identified since it deviated from the pre-defined *it is + adjective/verb participle that* structure. While Thematizer had greater difficulty with conflated coreference resolution in the training texts, it was the cleft pattern mismatches in the test dataset that proved more troublesome.

The final shared error classes between training and test datasets were modal themes that Thematizer overlooked. Modal themes in test texts were the only marked theme class whose parse yielded an accuracy less than 90.0%. An error frequency of 17.9% or 8 of 39 marked theme errors in the test dataset helps explain why modal themes suffered most in particular. Seven of the modal themes in the test texts were entirely missed due to Thematizer encountering novel modal themes, thereby compromising the resulting F_1 score. As for novel modal themes in the training texts, the three errors were due to Thematizer misparsing vocatives as the grammatical theme instead of a marked theme. In doing so, they were simply overlooked, which, for simplicity's sake and comparison purposes, were categorized here as a novel model theme. Modals that appeared within the training texts were used to populate the pre-defined look-up tables that Thematizer should look for. Theoretically, modal themes should have been accounted for in the training texts; however, if not identified in the index identification task due to dependency misparses, then modal, let alone marked, themes could not be classified either.

The remainder of the error classes summarized in Figure 5-6 are unique to each dataset. Lexical repetition (10.6%) and section header (8.3%) misparses affected training texts alone and were yet another byproduct of the selfsame errors during index identification. Both cases were present in circumstantial themes alone, whereby Thematizer would occasionally mark the first case of repeated lexis (e.g., *too* in *too, too many people*) as a circumstantial theme. This was due to erroneous dependency parses, whereby the first instance of the repeated lexeme was marked as dependent on the finite verbal root of the matrix clause. Instead, as in the example given, it should have been a dependent of the adverbial's head, i.e., *too* is dependent on *many*, which is dependent on *people*. Since the adverbial was denoted as dependent on the finite verbal root, Thematizer then misparsed it as a circumstantial theme.

Section headers, including numbered lists, were a similarly frequent error and were assumed to be a circumstantial theme due to appearing before the grammatical subject or due to ambiguity

in its part of speech. For instance, in the section header *3. Round up your purchases.*, there is no grammatical theme, which should have caused Thematizer to relegate the entire phrase to the rheme. Instead, Thematizer often extracted the number as a dependent sentence constituent and demarcated it as a marked theme. Since it is a number, it was then classified as a circumstantial theme. Otherwise, ambiguous syntactic classes resulted in Thematizer assuming the section header to be an independent sentence. This can be seen in the example header *Chewbacca Carries*, which is a compound noun and refers to the workout exercise of carrying equipment like the Star Wars character Chewbacca. Since *carries* is also the third-person singular form of the verb, the dependency parse identified it as the congruent verb of the grammatical subject *Chewbacca*. Such syntactic ambiguity, while rare, often manifested itself in misparses of circumstantial themes, although cases were also evident in grammatical themes during index identification.

The final error case applicable to hypotactic themes specifically was the occasional conflation of grammatical themes as non-finite relative clauses through gerunds or participle phrases. This occurred at a frequency of 5.3% or 19 of the 360 marked theme errors. An example of this error was the sentence *Forecasting a hurricane season challenges us to better understand how the atmosphere works*. There, *forecasting a hurricane* was marked as the hypotactic theme, *season* as the grammatical theme and the remainder the rheme. Instead, *forecasting a hurricane season* should have been the grammatical theme. On account of the intended grammatical theme functioning as a complex phrasal subject, Thematizer assigned such instances to the class of ADVCL (i.e., adverbial clause). This means that the adverbial clause, the hypotactic theme here, was dependent on the remainder of the clause and needed to be separated from the assumed matrix clause. As Thematizer then used the dependency ADVCL as a parameter in classifying a hypotactic theme as a non-finite relative clause, it defaulted to this marked theme class.

The misparses of the adverbial *so*, embedded themes and coordination fall under structural and modal theme errors and were unique to the training dataset. Individually, they comprised 19, 10 and 8 of the total theme errors, respectively. All three error cases, while separate classes, followed the same misparse characteristic: structural themes were incorrectly extracted from their bound constructions. With the adverbial *so*, for instance, Thematizer would occasionally extract it from its bound construction *so that*, as in *So that you can finish the pavement, I've hired some other contractors*. Similarly, modal and structural themes were removed from the relative clause they were embedded in (e.g., *The book, which was so interesting, sold out immediately*). Finally, coordinating conjunctions *and* or *but* found within coordinated hypotaxis or prepositional phrases, for example, were taken from their syntactically bound clausal unit (e.g., *When the weather cleared up and people arrived*). While infrequent, these misparses emerged due to errors in the dependency tagging that took place during pre-processing and index identification.

The final class of errors to discuss from Figure 5-6 is the remaining misparse from the test dataset, fronted elements. In total, only a single case occurred as a result of an incorrectly split t-unit. Here, *The people who only want the ball so no one else gets to have a ball are having as much fun* was split into two t-units between *so* and *no*. As a result, Thematizer misparsed the *no* as an interjection and thereby a modal theme. Yet again, the consequence of t-unit misparses manifested in the marked theme parses.

From this discussion on error class frequencies, findings reveal that both circumstantial and hypotactic themes were misparsed most frequently. Across both datasets, 151 circumstantial themes and 98 hypotactic themes were either misclassified or overlooked. The reasoning behind the greater number of errors for these two classes of marked themes specifically is due to the

syntactic complexity that either can potentially achieve. Whereas structural and modal themes are most commonly single-word adverbials with limited modifiability to their construction, circumstantial and hypotactic themes enjoy rich realizational patterns through embedded, subordinate and coordinating structures. Finally, circumstantial and hypotactic themes belong to the most frequent marked theme type after structural themes, totaling 27.0% and 17.0%, respectively (compared to 38.4% for structural, 12.4% for projecting and 5.2% for modal themes). Thus, it could be argued that a combination of both their higher frequency in texts and greater syntactic complexity was the reason for their higher number of parsing errors.

Despite such complexities, Thematizer's ability to capture circumstantial and hypotactic themes with an average F_1 score of 94.6% across all text types reveals two important facets: firstly, an increase in accuracy of 5.5% over previous approaches to thematic parsing; and secondly, the successful operationalization of these two marked themes in an automated approach (in addition to projecting and structural themes). In fact, average F_1 scores of at least 91.0% in both training and test datasets indicate how well Thematizer was able to extract and identify marked themes.

As for modal and projecting themes specifically, these experienced the widest range of accuracy during parsing, whereby three and two cases of sub-80.0% F_1 scores emerged, respectively. While novel modal themes may continue to affect parsing accuracy in texts that Thematizer was not trained upon, inclusion of new modal themes when encountered can readily mitigate this issue. Projecting themes, however, represent a more difficult case due to the coreference that Thematizer occasionally assumes. Ensuring that projecting themes are not conflated with coreferential structures proves to be the primary parsing hurdle that Thematizer will need to overcome.

Finally, the most frequent and accurate marked theme class to be parsed was that of structural themes. This was primarily due to 67.8% of the structural themes being realized as a coordinating *and* or *but* after coordinated t-units were split. Therefore, assuming Thematizer parsed t-units appropriately, the subsequent parse of the resulting structural themes was nearly a guarantee.

While caution should be taken where modal themes are concerned due to the lower F_1 score they achieved specifically, overall, Thematizer's parsing ability of marked themes regardless of text type suggests an enriching of the thematic theoretical model that formed the basis of Thematizer's functionality. In fact, the findings that Thematizer yielded through marked theme analyses facilitated greater insight into the use of marked themes in the register of text types, as summarized in Figure 5-7.

There, the relative frequency of the five marked theme classes is presented with respect to register, which is categorized as formal, semiformal and informal. Examples of formal-register text types are the Wikipedia articles, L1/L2 university texts, newspaper articles and the linguistics textbook. Semiformal text types were the editorial, obituary, professional blogs and the gaming article. Informal text types were lyrics, the short story, the comments texts and personal blogs.

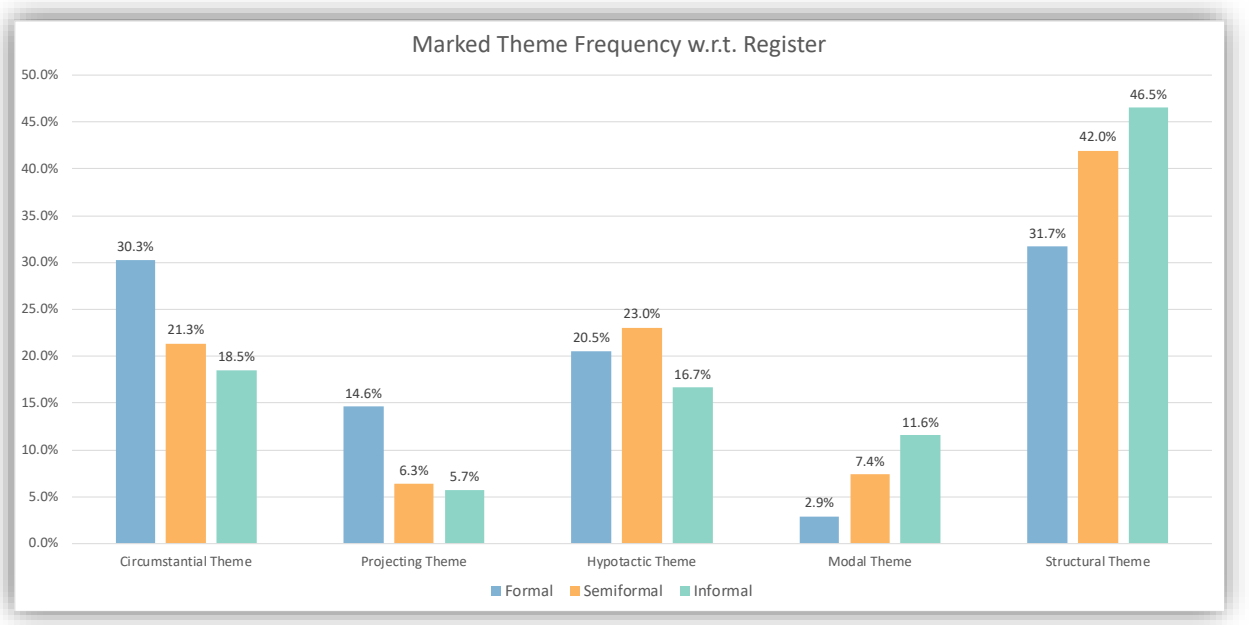


Figure 5-7: Relative frequency of each marked theme class with respect to register across training and test datasets. Formal-register text types were the Wikipedia articles, L1/L2 university texts, newspaper articles and the linguistics textbook. Semiformal text types were the editorial, obituary, professional blogs and the gaming article. Informal text types were lyrics, the short story, the comments sections (Reddit and gaming) and personal blogs.

From this figure, a strong trend can be seen in terms of the syntactic and semantic complexity of the marked theme class vis-à-vis register: the greater the formality of the text type, the greater the frequency of complex marked theme types. This is particularly evident in the decreasing frequency of circumstantial and projecting themes from formal to informal text types. Conversely, the less syntactically and semantically complex modal and structural themes exhibit an increasing frequency as the formality of the register decreases. The sole exception to this trend is that of hypotactic themes, whose greatest frequency was found in the semiformal register. However, even here, formal-register texts exhibited a higher frequency of hypotactic themes than informal-register text types.

This trend in marked theme frequency as a function of register sheds light on the texture characteristics of the grouped text types. Coordination through structural themes is much more frequent in less formal texts, whereas subordination and complex syntactic structures through circumstantial, projecting and hypotactic themes permeate more formal texts. The degree of marked theme complexity and frequency appears to be a reflection of the degree of complexity behind the subject matter presented in the texts. When more difficult content is discussed, more complex marked themes are used and at greater frequency. Conversely, subject matter that is either more straightforward or expressed to be more accessible to the audience enjoys greater use of simpler marked themes and structures.

Accessibility in terms of subjectivity versus objectivity is further reinforced through the frequency of modal versus projecting themes. Whereas the former imparts an interpersonal and emotive character to the text, the latter establishes distance and a more factual tone. As such, a higher frequency of modals in less formal texts, e.g., blogs and lyrics, is expected with a lower frequency in more formal texts, e.g., university essays and textbooks. The converse can then be stated for projecting themes, which are common in formal-register texts but rarer in informal-register texts. This finding is evident in the results for projecting and modal themes in Figure 5-7. The important note here is that marked themes' employment is relative to, but not the sole determining factor of, a text type's register, i.e., it is a cline. This allows for variations within

and across text types and marked themes to occur while reinforcing the mutual effect they have on one another.

On the basis of the results from Thematizer's second thematic parse, additional answers to the second research question can be formulated. The results from Thematizer's marked theme classification indicate the successful operationalization of thematic theory as it pertains to the identification and categorization of marked themes. This is substantiated by the high F_1 scores that Thematizer achieved and that exceeded the 89.1% gold standard. The sole exception to this was that of modal themes in test texts, whose F_1 score was 77.8%. Despite this single case, an increase in parsing accuracy across the board illustrates the strengths of Thematizer's syntactic and pattern-based approach to marked theme classification. This then reinforces the finding from the index identification task, whereby a solely syntactic parsing methodology resulted in insufficient operationalization of thematic theory. By complementing dependency-based parsing with pattern matching, Thematizer was able to more accurately capture marked themes. The insights into the relationship between marked theme use, syntactic and semantic complexity as well as register further reinforce the adroitness Thematizer has in analyzing and classifying marked themes.

5.4 Error Cases and Accuracy Rates for Thematic Progression Classification

The final set of results to discuss concerns itself with the third and final thematic parse that Thematizer performs, the classification of thematic progression patterns. As done in the previous sections, the F_1 scores that Thematizer achieved in this parsing task are first presented as a comparative summary of the accuracy for both training and test datasets. Afterwards, the specific error classes that affected the resulting parsing accuracy are outlined in addition to the reasons for the misparses that occurred. Key findings from these results close the discussion on Thematizer's thematic progression analyses.

To start, the goal behind this third parsing task was the classification of the text's thematic progression patterns. Included in that parse was the identification of the correct element(s) that instantiated thematic progression as well as the means of progression, e.g., lexical repetition or coreference. Therefore, misparses occurred if Thematizer classified the thematic progression pattern incorrectly (e.g., linear instead of continuous progression), if it misidentified the connecting element from the text for thematic progression instantiation or if it misidentified the means of progression. How well Thematizer was able to parse a text's thematic progression is summarized in Figure 5-8, which shows each thematic progression pattern and its F_1 score for the training and test datasets.

First of all, a decrease in F_1 scores and thereby parsing accuracy between training and test data is evident across most thematic progression patterns. On average, training data achieved an accuracy of $F_1 = 80.2\%$, whereas test data yielded an accuracy rating of 75.9% only: an overall decrease of 4.1%. Based on these F_1 scores, Thematizer was only sporadically able to exceed the 79.2% gold standard in the training texts, specifically with constant and gapped constant progression, macrotheme and thematic break patterns. Thematizer's failure to achieve the gold standard in all patterns but macrotheme progression illustrates its difficulty in classifying thematic progression in novel (test) texts compared to those it was trained on. What is more, Thematizer performed considerably worse with certain text types, such that the thematic analyses required considerable scrutiny in terms of output reliability.

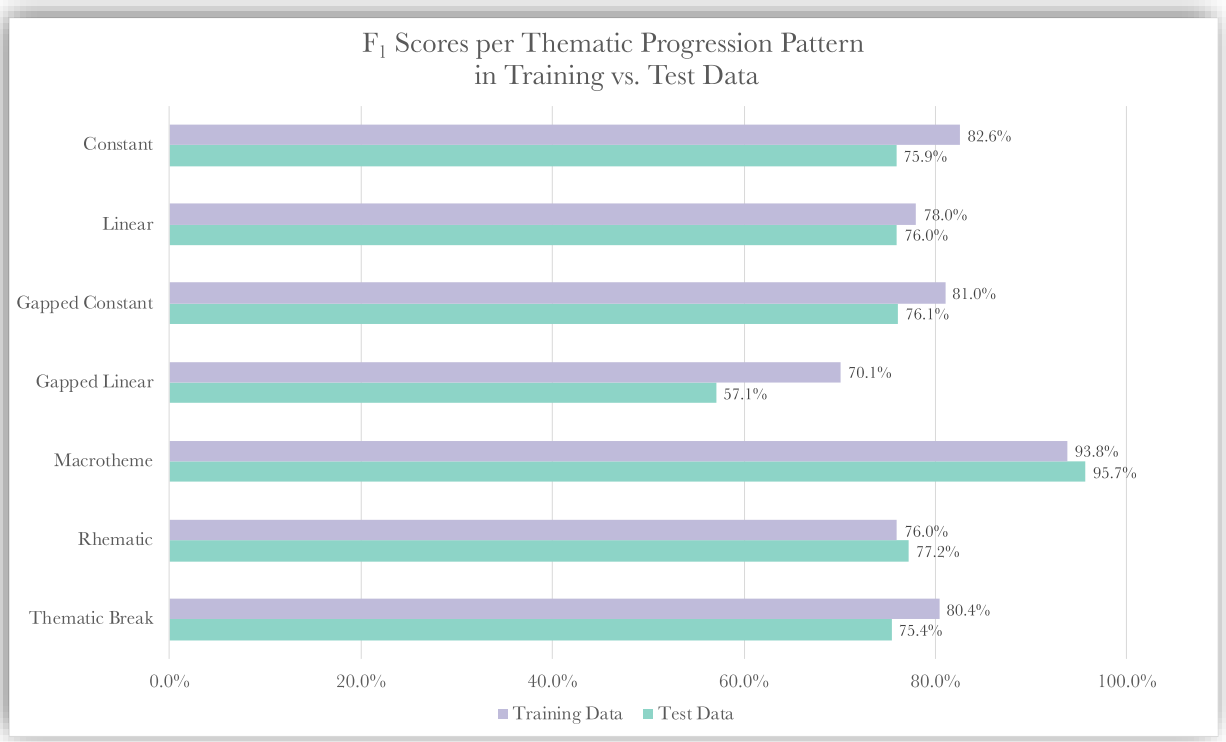


Figure 5-8: Accuracy rates Thematizer achieved across training and test datasets for each thematic progression pattern and in terms of F_1 score. Values above the gold standard of 79.2% indicate an increase in thematic progression accuracy over computational approaches from previous research.

The questionability of thematic progression results is reinforced by the 57.1% F_1 score that gapped linear progression achieved in the test dataset. In the training set, gapped linear progression also achieved the lowest F_1 score of all thematic progression patterns at 70.1%. This indicates Thematizer's tendency to overgeneralize gapped linear progression, either in seemingly defaulting to gapped linear over other progression patterns or in missing gapped progression where actually present.

On the other end of the F_1 score spectrum, macrothemes were correctly classified most frequently, achieving an F_1 score of 93.8% for the training data and 95.7% for the test data. Such high accuracy ratings are attributed to the pattern-based matching that Thematizer employs with the macrothemes identified through Latent Dirichlet Allocation. The identification of document-wide discourse topics as macrothemes therefore appears to be the soundest theoretical and programmatic approach to thematic progression analysis amongst the seven patterns.

Aside from gapped linear progression and macrotheme as polar extremes, the F_1 scores of the remaining progression patterns from the training data differed by 1.2% to 6.7% compared to the test data. A narrower range in these progression patterns' F_1 scores across datasets seems to indicate a degree of consistency in parsing as defined in Thematizer's theoretical and programmatic model for thematic progression classification. However, errors that emerged in the training dataset ultimately worsened in the test texts, as evidenced by the lower F_1 scores overall. The wider the F_1 score range between datasets, the greater the error classes permeated the particular progression pattern. This finding therefore suggests that particular aspects of Thematizer's parsing methodology for thematic progression were inadequate.

To help illustrate where Thematizer excelled in its thematic progression parse and where it failed, the heatmap in Figure 5-9 was tabulated using the F_1 scores for each progression pattern and text type analyzed (cf. Table A-2 and Table A-3 in Appendix A for a breakdown of their precision and recall scores). Average F_1 scores are provided as well for ease of comparison amongst the text types. Cells shaded in red to dark yellow indicate F_1 scores that were below the gold standard of $F_1 = 79.2\%$ and thereby indicate a comparative decrease in parsing accuracy. Light yellow to green cells conversely indicate an increase over previous computational approaches in terms of parsing accuracy.

Considering the thematic progression patterns' F_1 scores first, those for constant, gapped constant, gapped linear and thematic breaks stand out from the rest. In these four cases alone, sub-50.0% accuracies emerged, which partially explain the summarized accuracy ratings first introduced in Figure 5-8 above. Whereas gapped linear and thematic breaks experienced the greatest number of misparses, constant and gapped constant suffered egregiously in the Wikipedia article from the test data. In fact, all sub-50.0% F_1 scores were achieved in the test texts, which significantly hampered the accuracy of thematic progression classification.

The pairs of constant–linear and gapped constant–gapped linear exhibit a nearly identical trend in their F_1 scores: apart from the F_1 scores for the Wikipedia article and L1 university text in the test datasets, all F_1 scores for constant progression were higher than those of linear. Similarly, the F_1 scores for gapped constant were higher in every case than gapped linear's F_1 scores. While still somewhat variable depending on text type, Thematizer's greater accuracy with constant progression patterns suggests both the correct classification of this type and the overgeneralization of linear patterns. Gapped linear represents the exacerbated type of the two linear patterns, whose misparse frequency was the greatest.

Macrotheme and rhematic progression, while error-prone, were the only progression types to have achieved F_1 scores of 100.0% within multiple text types. That being said, rhematic progression still suffered from a wide accuracy span ranging from 56.1% at its worst for Wikipedia articles in the training dataset to 100.0% in four of the test texts. Therefore, despite the four perfect parses, rhematic progression was unable to achieve the same consistent rate of accuracy that macrotheme progression did across the board.

Turning to the text types specifically that were parsed with the greatest and least accuracy, the average F_1 scores in the final column of Figure 5-9 provide an initial answer to this accuracy span. The three text types that Thematizer classified worst all belonged to the test dataset and, specifically, were the Wikipedia article ($F_1 = 66.3\%$ and even 68.8% for the training dataset), the obituary ($F_1 = 66.8\%$) and the short story ($F_1 = 67.6\%$). The lower accuracy for the Wikipedia article is the result of the same errors that pervaded the Wikipedia articles from the training dataset, in fact at even greater frequencies. As touched upon in the following discussion, Wikipedia articles specifically suffered from Thematizer's failure to capture lexical entailment. This predominantly instantiated constant progression in Wikipedia articles but was assumed to be gapped linear on Thematizer's part. As such, both progression patterns were misparsed most frequently in this specific text type.

F_1 Scores of Thematic Progression Patterns w.r.t. Text Type									
Dataset	Text Type	Constant	Linear	Gapped Constant	Gapped Linear	Macrotheme	Rhematic	Thematic Break	Average F_1 Score
Training Texts	Wikipedia Articles	76.4%	70.4%	71.5%	58.5%	92.4%	56.1%	56.5%	68.8%
	L1 University Texts	83.3%	79.4%	79.7%	70.4%	91.7%	63.8%	67.5%	76.5%
	Blog Articles	85.6%	80.7%	85.4%	77.3%	97.5%	77.1%	85.0%	84.1%
	Lyrics	90.4%	79.4%	91.2%	81.3%	97.9%	94.1%	87.1%	88.8%
	L2 University Texts	89.3%	85.0%	86.2%	78.6%	93.2%	93.6%	63.2%	84.2%
Test Texts	Gaming News	90.9%	90.2%	88.9%	80.0%	85.7%	100.0%	66.7%	86.1%
	Newspaper Article	77.6%	66.7%	80.0%	58.8%	100.0%	100.0%	87.0%	81.4%
	Linguistics Textbook	80.0%	75.0%	85.7%	66.7%	100.0%	80.0%	50.0%	76.8%
	Reddit Comments	77.8%	76.5%	100.0%	88.9%	100.0%	80.0%	85.7%	87.0%
	Editorial	73.7%	69.2%	71.4%	60.0%	100.0%	100.0%	66.7%	77.3%
	Obituary	64.9%	63.6%	72.7%	50.0%	66.7%	100.0%	50.0%	66.8%
	Blog Comments	90.9%	77.8%	90.9%	50.0%	100.0%	75.0%	84.2%	81.3%
	Wikipedia Article	46.7%	79.2%	36.4%	36.4%	100.0%	80.0%	85.7%	66.3%
	L1 University Text	76.6%	83.3%	66.7%	57.1%	100.0%	60.9%	85.7%	75.8%
	Short Story	70.9%	63.4%	66.7%	46.2%	92.3%	85.7%	48.0%	67.6%

Legend	30.0%	50.0%	60.0%	80.0%	100.0%

Figure 5-9: Breakdown of F_1 scores that Thematizer achieved in classifying thematic progression patterns for each text type from both training and test datasets. Values below the gold standard of 79.2% are shaded in orange or red and indicate a decrease in parsing accuracy over computational approaches from previous research. Light yellow or green cells indicate an increase in parsing accuracy compared to the gold standard.

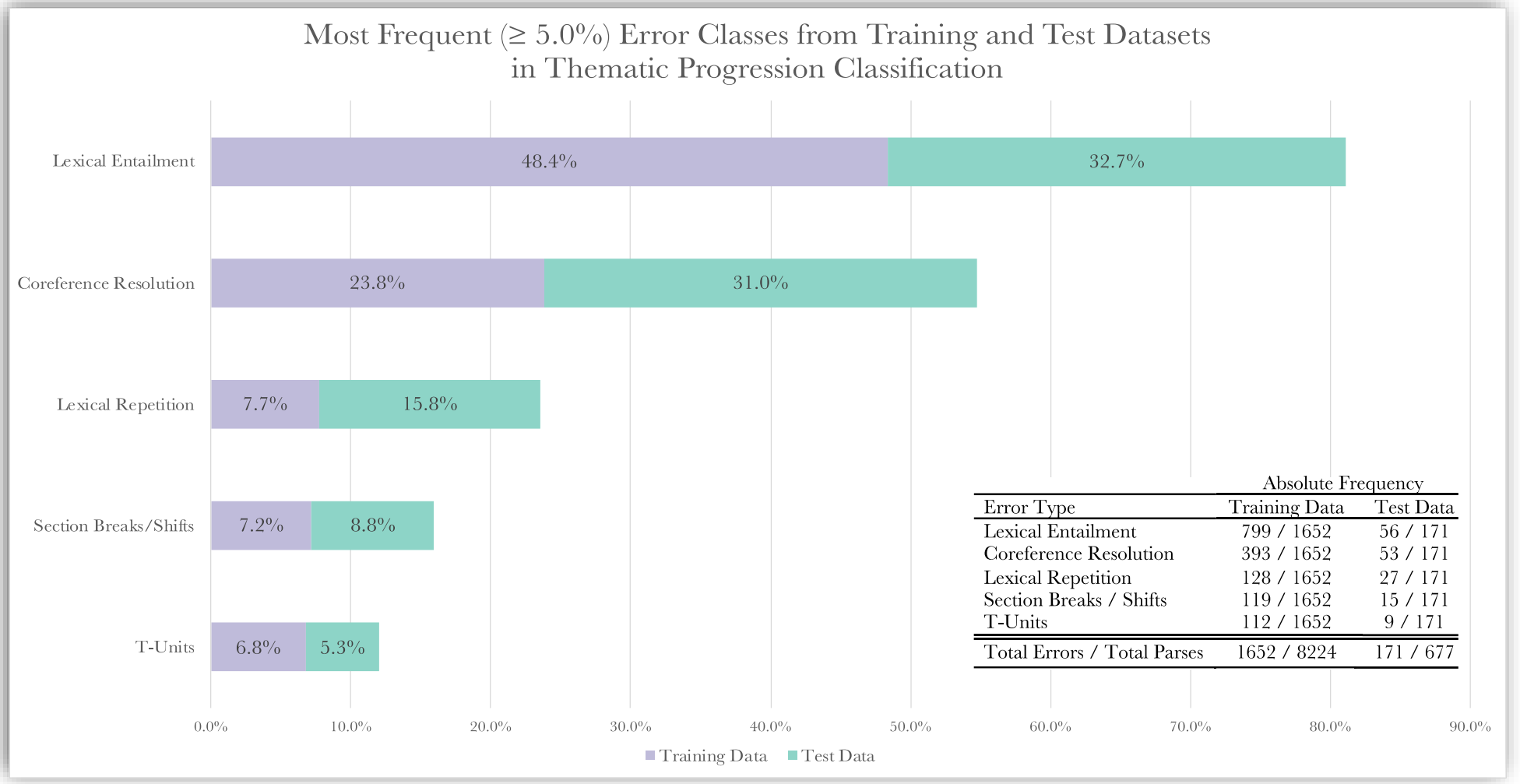
Due to coreference resolution misparses, an error class that proved problematic for all three of Thematizer's parsing tasks, the bibliographic nature of the obituary caused considerable difficulty for Thematizer. Unresolved coreference chains between proper nouns and antecedents were the ultimate cause behind the high number of misparses. This then caused misparses to ensue throughout all of the thematic progression patterns but affected gapped linear and thematic breaks most. It should also be noted that the 66.7% F_1 score for the obituary's macrotheme pattern represents a single false-positive of the two macrotheme cases from the entire text. Therefore, only one error occurred in Thematizer's parse of macrothemes in the obituary text. Still, this single case cannot be dismissed from its contribution to the obituary's resulting F_1 score.

The short story from the test dataset was the final text type to achieve an average F_1 score far below the 79.2% gold standard. Here, the sub-50.0% F_1 scores for gapped linear and thematic break patterns resulted in the short story's average F_1 score of 67.6%. Thematizer's failure to resolve coreference and identify lexical repetition caused the greatest number of misparses in the text's constant and linear progression as well as their gapped counterparts. Where unresolved, Thematizer erroneously assumed gapped linear progression to be present via lexical entailment. Where that failed as well, Thematizer incorrectly defaulted to a thematic break. This then caused the number of thematic break cases to become overinflated and overgeneralized, similar to gapped linear progression.

Indeed, these three text types with the lowest F_1 scores – the Wikipedia articles, obituary and short story – reflect a misparse commonality that came to affect nearly each progression pattern: Thematizer's failure to resolve coreference or lexical repetition most frequently led to gapped linear progression becoming the default progression pattern through assumed lexical entailment, followed closely by gapped constant as a default. Should either of these have failed, then Thematizer ultimately defaulted to a thematic break, which often became a catch-all class for progression patterns that were otherwise unaccounted for. Thus, the number of false-positives for this set of progression patterns – gapped constant, gapped linear and thematic break – became inflated in text types with high occurrences of coreference and lexical repetition.

The converse of this finding is equally true, whereby higher F_1 scores indicate Thematizer's success in identifying lexical repetition and resolving coreference. In such cases, thematic break's and/or gapped patterns' F_1 scores suffered considerably less, and the text (type) was able to achieve F_1 scores close to or beyond the 79.2% gold standard. This by and large explains how the remaining text types from Figure 5-9 were able to achieve average F_1 scores beyond the 75.0% mark in eleven of the fifteen text types. It should be noted, however, that lexical repetition and coreference resolution errors were not the sole causes behind lower F_1 scores. Yet where gapped patterns and thematic breaks are concerned, these errors played a critical role in the patterns' misparse.

All in all, 47 of all 105 thematic patterns (seven progression patterns across the fifteen text types) failed to exceed the 79.2% gold standard. This was a culmination of the highest number of parsing errors in Thematizer's three parsing tasks and is a clear indication of deficiencies in both parsing methodology and the operationalization of thematic theory. Overgeneralization of gapped progression and thematic breaks on account of falsely resolved lexical entailment, coreference resolution and lexical repetition appeared to be at the root of thematic progression misparses. However, as summarized in Figure 5-10, these were not the only error classes to have reduced Thematizer's parsing accuracy in its third parsing task.



	Absolute Frequency	
Error Type	Training Data	Test Data
Lexical Entailment	799 / 1652	56 / 171
Coreference Resolution	393 / 1652	53 / 171
Lexical Repetition	128 / 1652	27 / 171
Section Breaks / Shifts	119 / 1652	15 / 171
T-Units	112 / 1652	9 / 171
Total Errors / Total Parses	1652 / 8224	171 / 677

Figure 5-10: Relative and absolute frequency of error classes from Thematizer's third parsing task, thematic progression classification, across training and test datasets. Errors with a frequency less than 5.0% (thematic referent, missed macrotheme, interrogatives, deixis and theme-less sentences) are not included in the absolute frequency table tallies but constituted the remaining 101 misparses for the training data and eleven misparses for the test data.

Compared to the previous two parsing tasks, the error classes for thematic progression classification can be broken down to five core errors: lexical entailment, coreference resolution, lexical repetition, section breaks/shifts and t-units. These error classes occurred in both datasets at a minimum frequency of 5.0%. Minor errors that occurred less than 5.0% of the time and that are not summarized in Figure 5-10 were misparses concerning thematic referents, missed macrothemes, interrogatives, deixis and theme-less sentences. The total number of parsing errors (including those occurring with a frequency less than 5.0%) in the training dataset was 1652 of the 8224 parses (20.1%) compared to the 171 errors from the 677 parses in the test dataset (25.2%). From these relative frequencies, it is evident that test texts more often experienced the error classes summarized in Figure 5-10 than the training texts.

Lexical entailment was the most problematic error class for both datasets, although Thematizer had greater difficulty with the training datasets (48.4% for training vs. 32.7% for test datasets). The difference in error frequency between both text types shows an equitable difficulty in accurate cosine similarity parsing. The higher error frequency for training texts was due to their overall greater number of lexical entailment cases: altogether, 32.6% of the thematic progression patterns were developed via lexical entailment in the training datasets. Conversely, only 22.9% of the thematic patterns were from lexical entailment in the test datasets.

The underlying cause behind the misparses was the upper and lower bounds defined for each progression pattern (cf. Chapter 4.10.5). These were postulated in order to allow Thematizer to scrutinize the thematic progression patterns more closely on the basis of the returned cosine similarity values. The high number of parsing errors, however, show that these bounds proved too conservative for some patterns (non-gapped progression patterns) and too liberal for others (gapped progression patterns). This caused Thematizer to overgeneralize patterns as gapped and overlook either continuous or linear progression. The pervasiveness of lexical entailment errors from erroneous cosine similarity tests was ultimately a determining factor in the overall reduction in accuracy for Thematizer's parse of thematic progression classification.

The second most common error case from Figure 5-10 was that of coreference resolution. This appeared again in both datasets but affected the test dataset more readily (23.8% for the training vs. 31.0% for the test datasets). This error class may have become familiar by this point as it affected each and every parsing task to varying degrees. If originally found in the index identification parse, the same misparses impaired both marked theme and thematic progression classification. How the coreference misparses emerged was in one of two forms: either Coreferee incorrectly identified the antecedent in the previous sentence or Thematizer employed the incorrect coreference index and/or textual realization to resolve coreference. In either case, this caused Thematizer to overlook coreference where present or base the progression pattern on the incorrect sentence constituent that instantiated thematic progression.

The greater number of coreference misparses in the test dataset was due to the greater number of coreference cases that needed to be resolved in those texts. Despite being a common means of development in both datasets, coreference was found in 12.3% of the test texts and only 10.1% of the training texts. The obituary, Reddit comments, blog comments and gaming news article from the test texts, in particular, revealed frequent use of coreference and correspondingly high misparses. In contrast, the majority of the training texts had a more formal register, where the use of first-person pronouns was less. Where errors did emerge in formal-register texts, however, was with Thematizer's assumption of coreferentiality in projecting themes. In doing so, the dummy-*it* from the projecting theme was assumed to be the subject and grammatical theme that caused thematic progression across sentences. Here, the conflation of the dummy-*it*

with the coreferential pronoun *it* reinforces the inadequacy behind the pre-defined pattern for clefts that Thematizer used when searching for cases of clefts alone.

Lexical repetition formed the third most common error classes in Thematizer's third parsing task, again affecting test datasets (15.8%) more than training datasets (7.7%). If Thematizer was unable to identify individual lexemes or entire phrases that were repeated across concomitant sentences to instantiate thematic progression, then the parse was considered a failure. For example, if the phrase *poetic muse* was repeated in the themes of two sentences but Thematizer identified *muse* alone as the repeated lexeme, then it was marked as an error. The repetition of lexical items using a different part of speech also led to failure cases despite lemmatization of search terms. Failure to identify progression from *juxtaposition* in one sentence to *juxtaposing* in the gerund form in the next sentence was an instance of such a case.

Of these errors, it was partial identification of the repeated phrases that caused the greatest number of misparses. Noun chunks that Spacy returned and that Thematizer used as search tokens for lexical repetition were occasionally incomplete. Complex noun phrases in particular were returned inconsistently despite uniform syntactic realization patterns and dependencies in texts. Since lexical repetition was the most frequent means of progression in both datasets (49.8% for training texts and 41.9% for test texts), their parse had significant effects on the resulting accuracy.

Thematizer's failure to identify shifts in rhetoric or the introduction of new sections in text by means of thematic breaks formed the next common error class in thematic progression classification. Thematizer defaulted to thematic breaks if coreference resolution, lexical repetition, macrotheme instantiation or cosine similarity could not account for progression. Often, however, Thematizer misparsed lexical repetition or macrotheme instantiation, causing the test to fail and thematic break to be chosen unnecessarily. Simultaneously, erroneous cosine similarity values between theme and rheme elements prevented a thematic break from being reached despite being present. In other words, if semantic similarity fell within the upper and lower bounds of a particular progression pattern, then Thematizer erroneously marked the progression as such without consideration of a thematic break.

These misparses occurred in a total of 119 cases for the training texts and fifteen cases for the test texts. Although these amount to 8.8% of the errors or less, thematic breaks constitute an important structural component and rhetorical device in writing. As such, accurate and consistent identification of shifts in the rhetorical sections of a text to trace discourse development was defined as a critical component of Thematizer's functionality.

Misclassification of thematic progression patterns due to recurring misparsed t-units forms the final most frequent class of errors in the third parsing task (6.8% in training texts and 5.3% in test texts). Similar to coreference resolution cases, t-units that were incorrectly parsed in the index identification task led to misparses in the classification of thematic progression patterns. Misparses stemming from incorrectly split t-units led to the introduction of new sentences that were dependent in nature and were therefore marked as having no thematic progression. Here, Thematizer occasionally split independent clauses from their subordinate clauses after dependency tests returned a false-positive for independence. In other words, a grammatical subject and congruent verb were found in subordinate clauses that substantiated independence for Thematizer. Resolving t-unit misparses from the initial parsing task would thereby prevent their emergence in thematic progression classification.

Thematizer's poor accuracy in thematic progression classification gives rise to the finding that the operationalization of thematic theory remains deficient. Similar to the index identification task, the results shed doubt on the parsing methodology employed for certain thematic progression tests. It is therefore questionable at best whether cosine similarity together with a stipulation of upper and lower bounds for semantic similarity values was an appropriate means of resolving lexical entailment. While semantic tests via cosine similarity were one foundational reason for the reduced accuracy, syntactic tests via coreference resolution proved equally denigratory in the analytical output. Here, the way in which Thematizer indexically makes use of Coreferee's coreference indices and textual realizations must be scrutinized further on account of the variability in the parsing accuracy.

Where Thematizer succeeded best in thematic progression classification was with its pattern-based methods used for lexical repetition and macrotheme instantiation. Misparses from lexical repetition were indeed a common error class between both datasets. However, it is argued that tracing lexical repetition through pattern-based matching accounted for the increased minimum parsing accuracy. Continued use of pattern-based matching and finding a replacement for cosine similarity testing to resolve lexical entailment could then result in more reliable output for thematic progression classification.

Aside from determining Thematizer's parsing accuracy of thematic progression classification, the present work also explored the potential relationship between thematic progression patterns and text type. Previous researchers have postulated that certain text types are emblematic of particular thematic progression patterns as a reflection of the text's method of development and underlying discourse function (cf. Hasselgård 2020; Hawes & Thomas 1997b; Swales 1990; Berry 1995; Matthiessen 1995; Fries 1995; Martin 1992). In order to confirm or repudiate these findings, the frequency distributions of the thematic progression patterns from the training texts were used. Since only ten test texts of a single text type were used, the frequency distributions of their thematic progression patterns were not considered.

Testing for statistical significance between thematic progression pattern and text type was achieved through χ^2 (chi-squared) tests under the condition of a 0.05 significance level (α = 95.0%). Here, the null hypothesis of the χ^2 test was that no significant relationship exists between thematic progression pattern and text type. The results of these tests are summarized in Table 5-2, where the p-value is given for each thematic progression pattern with respect to text type. Values that proved significant are given in bold.

A statistical significance between thematic progression pattern and text type proved to be evident in all text types, albeit to varying degrees. While blog articles and lyrics were shown to have five and even six significant thematic progression patterns, respectively, the remaining text types exhibited significance with three or four patterns. Considering the thematic patterns individually, all but one (gapped linear progression) revealed a statistical significance with at least three and up to five of the text types.

Thematic Pattern	Wikipedia Article	L1 University Text	L2 University Text	Blog Article	Lyrics
Constant Continuous	p = 0.37	**p = 0.001**	p = 0.13	**p < 0.0001**	**p < 0.0001**
Gapped Continuous	p = 0.35	p = 0.41	**p = 0.016**	**p = 0.003**	**p < 0.0001**
Gapped Linear	p = 0.09	**p = 0.037**	p = 0.79	p = 0.17	p = 0.74
Linear	**p = 0.018**	p = 0.27	p = 0.98	**p < 0.001**	**p < 0.0001**
Macrotheme	**p < 0.001**	**p < 0.0001**	**p < 0.002**	p = 0.09	**p < 0.0001**
Rhematic	**p = 0.003**	p = 0.54	**p < 0.043**	**p < 0.0001**	**p < 0.0001**
Thematic Break	**p < 0.0001**	**p < 0.0001**	p = 0.62	**p < 0.0001**	**p < 0.0001**

Table 5-2: Resulting p-values from χ^2 tests on the relationship between thematic progression pattern and text type. The high number of statistically significant values (in bold) indicate that thematic progression pattern alone cannot serve as a marker for text type membership or identification with the potentially sole exception of gapped linear progression.

Taken together, these findings indicate that a single thematic pattern is not representative of a specific text type, with gapped linear progression being the sole exception for L1 university texts. This means that the prevalence of a particular thematic progression pattern alone cannot be used as a delineating texture characteristic for the identification of a text's membership to a specific text type. For instance, lyrics employ all thematic progression patterns but gapped linear progression to such a significant degree that they are all emblematic of a lyrical text. However, the frequency of these thematic progression patterns alone would be an insufficient determining factor for the categorization of a text as belong to the text type of lyrics. The same then applies to the other text types on account of at least three thematic progression patterns exhibiting statistical significance.

Again, the sole exception to this finding is that of gapped linear progression for L1 university texts. As gapped linear progression attained statistical significance with this single text type, it could be viewed as a potential candidate for a text's membership to the text type of L1 university texts. Gapped linear progression, also known as split rheme progression, has been found to be a representative thematic progression pattern in academic writing (McCabe 1999: 203; Jalilifar 2009: 98). The statistical significance found between gapped linear progression and L1 university texts therefore reinforces this finding. On a related note, the lack of significance between gapped linear progression and L2 university texts could indicate non-natives' tendency to prefer gapped constant progression, which proved significant. Non-natives appear to prefer developing the same thematic foundation across sentence clusters as is more common in narratives or texts with a narrative style (McCabe 1999: 203). In academic writing, however, a typical zig-zag structure emerges with gapped linear progression as a reflection of the presentation of propositional content in a cause-effect relationship. In the case of either native or non-native texts, however, gapped linear progression's statistical significance should still be taken with consideration since its simple (i.e., non-gapped) equivalent, linear progression, did not prove to be statistically significant. This is despite the same discursive function that linear progression shares with gapped linear progression.

The broader conclusion to be drawn from the findings in Table 5-2 is that thematic progression patterns should be considered an additional, not a solely defining, texture characteristic. Just as lexical density can reflect a text's register and contextual influences, it belongs to a greater

repertoire of the text type's characteristics that should not be reduced to a single parameter. Therefore, the significant relationships summarized above should be viewed as a strong tendency, but by no means a prescriptive metric, to be used in the composition of text.

In consideration of the results presented in this section, the following five core findings can be summarized: Firstly, Thematizer's accuracy suffered most in its final thematic parsing task, with test datasets achieving an even lower F_1 score than the training datasets. Thematizer's general failure to meet or exceed the 79.2% gold standard is a reflection of the recurring errors that permeated the parses. This, then, puts into question the reliability of Thematizer's thematic progression output. Secondly, gapped progression and thematic breaks proved to be overgeneralized patterns that became the default if previous thematic progression tests failed. Here, cosine similarity values, coreference misparses and misidentification of lexical repetition contributed most to failure cases. Thirdly, cosine similarity values together with upper and lower bounds frequently resulted in failing to account for lexical entailment as a means of progression. This formed the predominant error case in thematic progression classification and thereby contributed most to reducing Thematizer's overall accuracy. Fourthly, the results presented here indicate that Thematizer was unable to attain consistent operationalization of thematic structure for thematic progression. This can predominately be attributed to the insufficient semantic testing methodology and coreference resolution. Fifthly and finally, due to numerous thematic progression patterns revealing a statistically significant relationship with more than one text type, they cannot be viewed as a definitive texture characteristic for ascribing text type membership.

5.5 Summary of Key Results from Thematizer's Parsing Functionality

The present chapter set out to present the results from the three thematic parsing tasks that Thematizer performed in its text analyses: index identification, marked theme classification and thematic progression classification. Through F_1, precision and recall scores, the accuracy yielded for each of these parsing tasks was quantified as the degree of Thematizer's parsing success or failure. The treatment of predominant and recurring errors that occurred in each of the parses helped to shed light on how they affected the resulting parses and their accuracy scores. With these key results, answers to the present work's research question of Thematizer's operationalization of thematic theory were able to be formulated.

Parsing Task	Training Data F_1 Score	Test Data F_1 Score	Gold Standard F_1 Score
Theme/Rheme Index Identification	**85.8%**	92.0%	89.1%
Marked Theme Classification	94.9%	93.4%	89.1%
Thematic Progression Classification	80.2%	**75.9%**	79.2%
Thematizer's Final F_1 Score	85.7%	85.4%	79.2%

Table 5-3: Individual and overall F_1 scores for each of Thematizer's three parsing tasks compared to the gold standard from previous research and software for computational approaches to thematic analysis.

Beginning with a recap of the F_1 scores Thematizer achieved, the training dataset proved slightly more accurate in its parses than the test dataset, as shown in Table 5-3 (repeated from Chapter 5.1). The decrease in the test dataset's accuracy was the result of the high number of misparses in the marked theme and thematic progression classification tasks. This was in spite of the considerable increase in parsing accuracy for the index identification task. Whereas marked theme classification alone was able to exceed the gold standard, the other two parsing tasks achieved mixed results. For the index identification task, only the test dataset exceeded the gold standard; for thematic progression classification, it was the training test alone that surpassed the gold standard.

The reason for an overall reduction in parsing accuracy across text types could be traced back to error classes both unique to and shared amongst the three parsing tasks. Certain errors that emerged in the index identification task had a cascading effect whereby they impacted the subsequent parsing of marked themes and thematic progression patterns. By and large, such error cases were the misidentification of the subject or verbal root index, t-unit misparses, coreference resolution and, to a degree, marked theme indices. Cleft pattern mismatches, lexical repetition and shifts in rhetorical sections of the text formed the secondary group of common error classes shared amongst multiple parsing tasks.

On a task-by-task basis, misidentification of the subject or verbal root index affected Thematizer's parse most in the index identification task. For marked theme classification, t-units that were either erroneously split or remained unsplit caused the greatest number of errors. Next, the majority of errors during thematic progression classification stemmed from Thematizer's failure to resolve lexical entailment or coreference. Finally, the remaining errors to occur were less pervasive and largely or entirely unique to the parsing task in which they emerged. Their continued and collective occurrence contributed to a marked reduction in the parsing accuracy of both datasets overall. How and why these individual errors occurred will be the topic of discussion in Chapter 6.

With the help of the resulting F_1 scores and error class frequencies, the degree to which Thematizer successfully operationalized thematic theory was ascertained. This formed one of the two research questions that motivated the present work and is considered to be a quantifiable indication of how well Thematizer was able to make thematic structure accessible to writers. Ultimately, only marked theme classification was shown to have successfully operationalized thematic structure on account of its F_1 scores exceeding the gold standard. The frequency of errors in index identification and even more so in thematic progression not only caused Thematizer's resulting accuracy to suffer; it also made the reliability of the parsing results too variable and therefore partially unreliable. For index identification, dependency misparses and the lack of corresponding pattern-based parsing for the identification of thematic constituent spans readily prevented complete operationalization. For thematic progression, inconsistent operationalization was attributed to the high variance in Thematizer's output with respect to accurate classification of each thematic progression pattern. The high number of lexical entailment and coreference errors, specifically, substantiate this finding.

The results presented in this chapter not only provided answers to the second research question of this work but also gave rise to initial key findings regarding Thematizer as a tool for automated thematic analysis. The performative results shed light on how well thematic structure was able to be traced by computational means and, equally as important, the key errors that caused failures. With an understanding of what caused Thematizer's parses to fail in each parsing task, the next chapter will outline how and why those errors came to be. This will provide the necessary detail for drawing final conclusions about the operationalization of thematic theory and the programmatic methodologies that underlie Thematizer.

Chapter 6 – Discussion of Key Error Classes and Results

This penultimate chapter addresses the key findings formulated in Chapter 5 together with the error classes that informed these findings. The purpose of this chapter's discussion is an in-depth analysis of the causes behind the error classes in Thematizer's three parsing tasks. An examination of the parsing errors' emergence and recurrence will facilitate amendments to the key findings from Chapter 5 such that final conclusions about Thematizer, its parsing functionality and operationalization of thematic theory can be drawn. How these findings and conclusions refute, confirm or build upon previous research will also form a central point of discussion, where applicable. Finally, Thematizer's areas of application and potential use cases will be outlined while summarizing the conclusions drawn from the present work.

Chapter 6 is broken down into four core sections. Chapters 6.1 to 6.3 examine the error classes that emerged in the thematic parsing tasks of index identification, marked theme classification and thematic progression classification. In each of these sections, the relevant error classes identified in Chapter 5 are reintroduced and treated individually. Treatment involves an explanation of what the error specifically was, how it manifested in the parse, and the programmatic deficiencies that caused the error. Then, the key findings pertaining to each parsing task are further refined and expanded with respect to the programmatic error classes. Chapter 6.4 then re-examines the two research questions that drove the research and development behind Thematizer.

6.1 Key Findings and Error Classes: Index Identification

The goal of the present section is first and foremost to outline how the error classes for index identification emerged from a programmatic and theoretical perspective. This will illuminate the underlying deficiencies in Thematizer's parsing functionality and facilitate conclusions on the present work's approach to automating thematic analysis. The understanding gained through a treatment of the underlying causes will then come into play in Chapter 6.1.1, where they will be further contextualized with the key findings from Chapter 5.

Thematizer's first parsing task, index identification, was shown to be the second most accurate thematic parse, achieving $F_1 = 85.8\%$ for the training texts and $F_1 = 92.0\%$ for the test texts. Across both datasets, dependency errors formed the foundation of the misparses and were instigated by comma (mis)placement, clause dependency errors, disambiguation and coreference resolution. These, in turn, were made manifest through the specific error classes that proved most problematic during index identification: errors concerning subject/root index identification, marked theme spans, t-units, cleft pattern mismatches, direct quotations, section headers and fronted elements. Note that these errors presented in Figure 6-1 may not be discussed in the same order they appear. Instead, they will be grouped according to the underlying parsing issue that caused their emergence. This will allow for an individual treatment of each error class while grouping them into shared underlying causes.

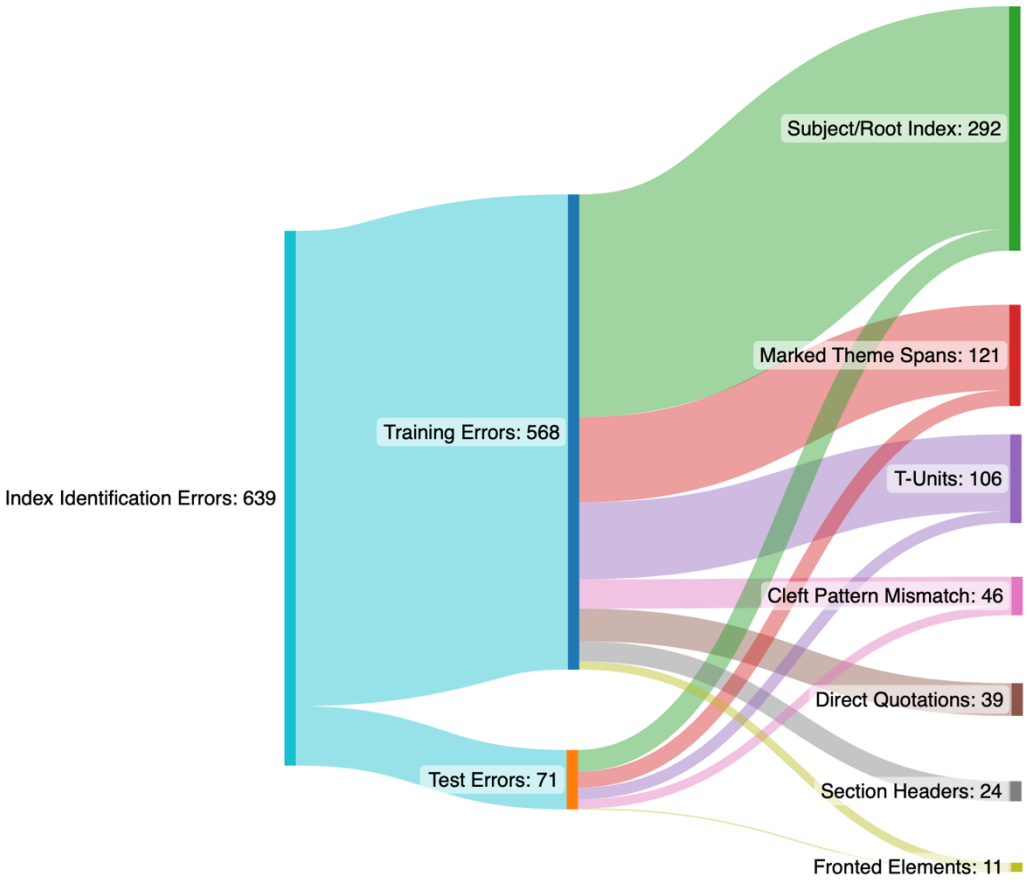


Figure 6-1: Summary of error classes in both training and test datasets for the index identification task. Errors are given as absolute frequencies.

Index Identification Errors due to Commas and Ungrammaticality

The first underlying cause was that of dependency misparses on account of ungrammaticality and misplaced or missing commas. This ultimately affected how subject/root indices and marked themes were parsed and are illustrated in relevant examples to follow. The text in bold is what Thematizer erroneously returned as the grammatical theme as indicated by the asterisk. Where the grammatical theme should have ended is indicated by the || marker.

(1) ***Accents or dialects** || **immediately tell one where a person is from for each region in every country and even down to individual cities in England's case, has it's** own dialect.

While *accents and dialects* alone constitute the grammatical theme in (1), Thematizer extended the theme span beyond the matrix clause's finite verb *tell* and adverbial dependent *immediately*, where the rheme span should have started. Here, the reason for the misparse can be traced back to the ungrammaticality of the sentence and misplacement of the comma. Firstly, the finite verb *has* towards the end of the sentence has no congruent subject, i.e., there is no singular subject that is syntactically dependent on this verb. Secondly, the contraction *it's* should have been in possessive form. However, even in the possessive, it is unclear what its referent should be although *region* or *country* could be potential candidates. Finally, the placement of the comma in front of the phrase *has it's own dialect* caused the entire clause beforehand to become dependent, functioning similar to a subordinate clause, albeit incorrectly.

Since dependency parses rely on grammatically correct sentences for accurate analysis, ungrammatical formulations such as (1) invariably produce incorrect output. Thematizer must then make use of dependency parses regardless of the input. Where grammatical inaccuracies occurred, they manifested most commonly, but not solely, in grammatical theme spans. This is a prime example of the computer science saying, "garbage in, garbage out," and cannot be avoided in parses unless grammatical issues are resolved beforehand.

(2) ***Several species of honeyeater, including the white-naped, yellow-faced, New Holland, and occasionally white-plumed and crescent honeyeaters, and** Eastern spinebills[,] have been observed foraging.

(3) ***And, if the undernutrition is moderate** enough[,] then these maternal metabolic actions compensate adequately for the fetus.

Compared to misparses stemming from ungrammaticality, the misplacement or absence of commas as in (2) and (3) contributed considerably more to grammatical theme misparses (correct comma placement is indicated by the [,] marker). In (2), the grammatical theme *several species of honeyeater* is augmented with a trailing reduced relative clause instantiated by *including*. The relative clause ends after *Eastern spinebills*, which Thematizer marked as belonging to the rheme but should have been included in the grammatical theme span. This would have been the case had the author included the required comma after *Eastern spinebills*. The same error is evident in (3), where a comma should have been placed after *enough* to mark the end of the *if*-clause. Its absence then resulted in an erroneous dependency parse since the beginning of the grammatical theme could not be identified correctly.

Correct use of a comma to offset a subordinate clause is pivotal in Spacy's ability to establish the correct grammatical dependencies in a given sentence. When subordinate clauses appear together with an independent matrix sentence, two sets of subjects and congruent verbs appear: one set for the subordinate clause and another for the matrix sentence. For example, in (3), the subject-verb set in the subordinate *if*-clause is *the undernutrition + is*, and the set in the matrix clause is *these maternal metabolic actions + compensate […] for*. In order to distinguish between both sets, Spacy marks the subject-verb set of the subordinate clause with an XCOMP or CCOMP dependency (Nivre 2022). These equate to clausal complements of a verb or adjective and are thereby automatically parsed as dependent. The resulting dependency for the matrix' subject-verb set is SUBJ and ROOT, whose indices Thematizer uses to demarcate the boundary between the grammatical theme and rheme. If a comma is missing, however, then this order may erroneously be reversed, or the ROOT of the sentence may be misidentified. In (2) and (3), it was the latter case that affected their parse. Missing commas can therefore result in grammatical theme spans being misidentified, but as shown in (3), it may also cause marked themes to be overlooked.

Conversely, misplaced or missing commas can cause Thematizer to parse marked themes alone while entirely missing the grammatical theme. This can be seen in (4), where the actual grammatical theme *outlines* was included in the circumstantial theme span.

(4)	*In the research project and even the documentation[,]*	*outlines*	*were missing.*
Incorrect Parse	*CIRCUMSTANTIAL THEME		RHEME
Correct Parse	CIRCUMSTANTIAL THEME	GRAMMATICAL THEME	RHEME

Due to the missing comma between *documentation* and *outlines* (as indicated again by the [,] marker), these two lexemes were assumed to form the noun phrase *documentation outlines*. This was then marked as the head of the prepositional phrase realized as a marked circumstantial theme. Here, the absence of the comma between *documentation* and *outlines* as a boundary marker between dependent and independent clauses again resulted in two misparses. While Thematizer falsely returned a grammatical theme alone in (3), it returned a marked theme only in (4) without a grammatical theme. The intimate connection that grammatical theme and marked theme spans have in their identification is such that the misparse of one often causes the misparse of the other.

Further examples of misparsed marked theme spans can be found in the misparse of fronted elements and lexical repetition during index identification. This problem was largely unique to lyrics, which enjoyed frequent use of vocatives, interjections and repeated lexis for emphatic, rhythmic or lyrical purposes. Examples (5) and (6) illustrate how commas caused fronted and repeated elements to occasionally be misparsed as grammatical themes or partially extracted marked themes.

(5)	*My baby,*	*just-a wrote me a letter.*
Incorrect Parse	?MODAL THEME	RHEME
Correct Parse	GRAMMATICAL THEME	RHEME

(6)	*Bad,*	*bad medicine*	*is what I want.*
Incorrect Parse	*CIRCUMSTANTIAL THEME	*GRAMMATICAL THEME	RHEME
Correct Parse	GRAMMATICAL THEME		RHEME

In (5), the grammatical subject *my baby* is marked as a modal theme after being identified as a vocative through the dependency parse. This is signaled by the use of a clause-separating comma and causes the resulting parse to lack a grammatical theme. Removal of the comma, however, would ameliorate this issue. It should be noted, however, that this is a fuzzy case: it could be argued that the intended or implied subject was *my baby*, which was elided to avoid repetition as a vocative, sentence-initial modal. Elided subjects were commonplace in the lyrics analyzed; however, as Thematizer was not always uniform in its parse with fronted elements followed by commas, the example was included for illustration of the potential error case.

Where Thematizer's parse was more overtly incorrect was the delineation of the first instance of *bad* in (6) as a circumstantial theme. Instead of marking the first *bad* as dependent on *medicine*, this token was interpreted as dependent on the finite verbal root *is*. Again, the comma here initiated this dependency: *bad* was interpreted as an adverbial of MANNER, indicating how something was done, similar to how *warily* is dependent on *walked* in *Warily, I walked the streets*. The colloquial form of the adverbial *bad(ly)* was what the parser then assumed. Since the token was separated from its noun head, it was considered an adjunct and not marked as belonging to the grammatical theme. It should be stressed that this only happened in fronted elements with repeated lexis. If unique compound adjectives described a noun, e.g., *the big, bad wolf*, then the dependency parse was returned correctly despite the comma. Regardless, this

error indicates the occasional tendency for dependency parses to conflate repeated lexis separated from its syntactic head by commas with MANNER adjuncts.

This reinforces the importance behind commas' functional use as a clausal boundary marker, as previous research has shown (Søgaard et al. 2018: 28-29). While humans may be able to determine or infer clausal and syntactic boundaries without commas, Thematizer relies on the dependency parses returned that are invariably affected by the presence or absence of commas. Introducing new and erroneous subordinate clauses, shifting tokens' dependencies to incorrect syntactic heads and producing false part-of-speech tags are a few of the effects that incorrect comma usage can have on dependency parsing. The correct use of commas as a syntactic factor in dependency parsing is therefore an indispensable determining factor behind Thematizer's parsing accuracy of grammatical and marked themes. While research has been done on the correction and automatic insertion of commas for improved dependency parses, results remain highly variable in terms of accuracy (e.g., see Israel et al. 2012 and Huang & Zweig 2002). Incorporating such models into Thematizer's dependency parses with respect to comma usage therefore poses a potential avenue for parsing improvements.

Index Identification Errors due to Clause Dependencies

The next major cause behind parsing errors was due to clause dependency misparses, which affected t-units and direct quotations primarily. In each of these cases, Thematizer was tasked with splitting two concomitant independent clauses joined via a semicolon, colon, hyphen, conjunctive adverbial, coordinating conjunction or quotation marks. As a reminder, since t-units possess their own thematic structure through the presence of a grammatical subject and finite verb, they are split into separate clauses (cf. Chapter 4.7.1). Examples of unsplit t-units and quotations can be found in (7) – (9), whose incorrect themes are, again, provided in bold. For comparison purposes, a correct t-unit that was not split on account of clause dependency is provided in (10).

(7) ***There are three limitations to take into account: The maximum aggregate** cannot exceed one-fifth of the dimension between forms.

(8) ***It [the seed starting mix] reads: "We** package our mix dry to avoid using plastic bags."

(9) ***For the first time I didn't know what the teacher wanted, so I** had to rely on my own creativity.

(10) **Prior to the past couple of decades, very few Devonian tetrapod taxa** were known: mainly Ichthyostega and Acanthostega, both from the uppermost Famennian.

A grammatical subject and congruent verbal root are present in both independent clauses separated by a colon in (7), a colon and quotation marks in (8) and by the coordinating conjunction *so* preceded by a comma in (9). Fulfillment of these conditions should have thus caused Thematizer to split them into separate t-units. Since the clause after the colon in (10) was dependent, Thematizer parsed the sentence correctly by not splitting the post-colon dependent clause from the independent clause.

The reason for the failure to split was because Thematizer was unable to identify both a subject and congruent root in both independent sentences. In (7), for instance, no grammatical subject (NSUBJ) index was returned since it starts with an existential, functioning as an ersatz subject

dependency. While Thematizer is equipped with identifying existential sentences as independent clauses, if exact dependencies delivered by Spacy were not met, then the parsing test failed. This is evident in (8), which contains two unmarked independent clauses. The unsplit t-unit here as well indicates Thematizer's failure to identify the grammatical subjects and congruent verbs on occasion.

Unsplit t-units are easily identifiable in Thematizer's output as the grammatical theme erroneously spans multiple sentences, as shown in the bold themes from (7) – (9). Upon closer examination, Thematizer was able to demarcate the correct boundary between the grammatical theme and rheme in the second independent sentence of each example. The error, however, is that the entirety of the first independent clause was included in the same theme.

The reason for this overgeneralization of the (grammatical) theme in unsplit t-units stems from the same error already discovered above with subordinate clauses in (2) and (3): clausal complement dependencies as XCOMP (predicative or clausal complement) and CCOMP (clausal complement). The dependency parsed that Spacy returned assumed that the first independent clause was actually dependent on the second independent sentence. When that happened, the finite verb of the first independent clause was marked as XCOMP or CCOMP, neither of which was the required root dependency that Thematizer needed to mark the boundary between grammatical theme and rheme spans. The root was identified in the second independent clause, however. Since the first independent clause could not be classified as a marked theme on account of its syntactic structure, Thematizer marked it incorrectly as the grammatical theme, up to and including the grammatical subject of the second independent clause. This, then, explains why the resulting grammatical themes in the examples above were so long.

Unsplit t-units were not the only cases of erroneous dependency errors for the class of t-units. Just as errors occasionally arose from Thematizer's failure to split two independent clauses, misparses from unnecessarily split t-units were also evident. In (11), the result of Thematizer splitting a dependent clause from its independent clause with themes in bold is shown.

(11) **I** became acutely aware of the size and quality of the shoulder. ***Or whether it** even existed, in some places.

Originally, a hyphen separated the second dependent clause from the independent clause, which initiated t-unit parsing. In this instance, Thematizer found a grammatical subject through *it* and a congruent verb through *existed*. This prompted Thematizer to split the clause unnecessarily from its independent matrix clause. What Thematizer failed to account for was the subordinating conjunction *whether*. This indicates a deficiency in the testing parameters for t-unit parsing, such that testing for a grammatical subject and congruent verb alone in both clauses may result in false positives. Instead, an additional testing parameter that checks for subordinating conjunctions or adverbials should also have been put in place. Otherwise, the output Thematizer delivers becomes a newly introduced fragment. This falsifies the marked theme and thematic progression classification tasks later since new t-units that should not have been separated are given their own additional thematic parsing.

The final case of clause dependency misparses took the form of independent clauses concatenated by commas, which also fell under the category of ungrammaticality. Instead of employing a period or semicolon, the author of (12) joined the two independent clauses with a comma, which caused a run-on.

(12) ***Three pig genes that trigger attacks from the human immune system were knocked out, six human genes that help to accept the donor organ** were added**.**

Again, the exceptionally long grammatical theme is due to *were* alone being identified as the verbal root and Thematizer allocating all sentence constituents up to that point to the grammatical theme. The finite verb of the first independent sentence was then marked as dependent through the finite verb's XCOMP dependency. This example is of particular difficulty, less so because of the clausal dependency and more so because of the use of a comma to join both clauses. Thematizer is not able to account for two concomitant independent clauses connected by a comma due to the considerable parsing overhead it would require: whenever Thematizer encountered a comma in a sentence – whether in listed constituents or as a clausal boundary marker – it would have had to test for independence. This would have added a significant amount of parsing time and, on the basis of the t-unit misparses discussed above, would have likely introduced even more errors. For that reason, it was decided to forgo separating t-units as in (12) despite the inevitable emergence of errors. As these errors occurred in less than 1.0% of the parses, the decision for processing efficiency and speed over splitting run-ons was made.

In spite of such errors occurring, users can interpret Thematizer's resulting output of grammatical themes spanning multiple independent clauses as an indication of potentially poor structuring, at least in formal-register texts. Since independent clauses are conventionally separated by a semicolon, colon or period, users can scrutinize how they wrote the specific sentence if Thematizer marks the grammatical theme up to the finite verb of the second independent clause. This can then prompt users to either maintain the run-on for stylistic purposes or to replace the punctuation mark connecting both independent clauses with a period to adhere to conventions for formal writing in their text.

Index Identification Errors due to Syntactic and Semantic Ambiguities

Dependency misparses stemming from syntactic and semantic ambiguities formed the next group of causes to affect section headers and certain marked theme spans. An initial form of syntactic ambiguity was already present in (4) above with *In […] document outlines*, which lacked a comma between *document* and *outlines* to mark the end of the prepositional phrase. This was also present in headers, which either lacked grammatical subjects or finite verbs and therefore should have been marked as rhematic only. This can be seen in (13), where the assumed grammatical subject *Jedi High* was marked as the grammatical theme. Since *Part 2:* appeared in front of the grammatical subject, it was incorrectly marked as the circumstantial theme. Instead, *Jedi High Jumps* functioned as a compound noun describing a form of Star Wars-inspired exercise to perform.

(13)	*Part 2:*	*Jedi High*	*Jumps.*
Incorrect Parse	*CIRCUMSTANTIAL THEME	*GRAMMATICAL THEME	RHEME
Correct Parse	RHEME		

The reason for the emergence of such errors as in (13) is both because of the subject-verb congruence found for the grammatical theme and the fronted phrase *Part 2* appearing before

the subject. However, as no finite verb actually exists in the clause, the entirety should have been given rheme status. The ambiguity that Thematizer failed to resolve was that *jumps* can belong to multiple syntactic classes: either a finite verb or the plural of the noun *jump*. As dependency parsers are trained to find verbal roots where possible – the root being an unequivocal requirement for all clauses of any length and composition – such parsers allocate root dependency to elements that possess this status with the greatest likelihood. Since *jumps* resolved the verbal root requirement, it was marked as such in the resulting parse, erroneous though it may have been. This was particularly problematic for section headers, which often elided finite verbs in favor of compound nouns or employed the participle of the verb. Invariably, the dependency parse returned the incorrect root, which then resulted in span misparses as shown in (13).

The misparse of *Part 2:* as the circumstantial theme was then a secondary effect of misidentification of the root. Upon being identified as a sentence-initial constituent in front of the grammatical subject and/or verbal root, section header titles such as *Part 2:* were often assumed to be a marked theme. Since the number two appeared in this case, Thematizer equated it to a circumstantial theme indicating TEMPORALITY. In fact, where no grammatical subject was identified in the section header, Thematizer often assumed the number in enumerated lists to be the grammatical theme. For example, in *5. Offer a welcome drink*, instead of parsing *offer* as an imperative, it was marked as the congruent verb of the number *5.* This again illustrates the dependency parse's tendency towards finding subject-verb pairs even when absent in the text.

Whereas syntactically ambiguous phrases were commonly misparsed as grammatical themes, semantically ambiguous noun phrases led Thematizer to assume their marked theme membership. In (14), Thematizer marked the noun phrase *intentionally misleading titles* as a hypotactic marked theme, which resulted in the grammatical theme being missed.

(14)	*Intentionally misleading titles*	*vaporize their credibility.*
Incorrect Parse	*HYPOTACTIC THEME	RHEME
Correct Parse	GRAMMATICAL THEME	RHEME

In the dependency parse, the noun *titles* was assumed to be the object of *misleading*, which indicates a right dependency, i.e., Spacy interpreted this sentence not as *the titles that intentionally mislead (people)* but *the titles that are intentionally misled*. Hence, Spacy assumed it was the titles which were being misled, not the titles that mislead readers. Since *titles* was given direct object dependency status, it could not be marked as the grammatical theme, which requires a grammatical subject dependency. In the remainder of the sentence, no other constituents could become the grammatical subject either due to the transitivity of *vaporize* causing *credibility* to be another direct object. This sentence parse therefore returned without a grammatical subject and, by extension, without a grammatical theme.

Ultimately, the parsing error stemmed from the ambiguity of *misleading*. This term is bidirectional in that it can entail the entity being misled or the one who misleads. Since Spacy works on the basis of probability functions to output a dependency parse from contextual and cotextual factors, it likely determined *titles* to be the entity being misled. That, in turn, prompted Spacy to allocate it direct object dependency, false as it may have been. Because Thematizer does not perform any semantic tests for dependency parses but instead relies on Spacy's output, the accuracy behind disambiguation parses ultimately depends on Spacy's syntactic analyses.

Index Identification Errors due to Clefts

The final class of errors to fall under dependency misparses is clefts, which were often conflated with coreferential structures. The introductory structure of non-projecting clefts through *it is* caused considerable confusion for Thematizer's pre-defined cleft pattern as it often assumed the dummy-*it* to be coreferential. Therefore, the *it* alone in a cleft structure was marked as the grammatical theme. The converse occurred equally frequently and is illustrated in (16) as a counterexample to non-coreferential dummy-*it* clefts shown in (15). An additional example of a misparsed projecting theme is then given in (17). Erroneous theme spans are shown in bold. Where the theme should have ended is given by the || marker.

(15) ***It** is || important to read. [Non-projecting cleft]
(16) ***It [the phenomenon]** || **was** researched thoroughly. [Coreferential *it*]
(17) ***It** should be questioned whether the claims made || were true. [Projecting theme]

In both (15) and (17), Thematizer assumed coreferentiality, i.e., it assumed the dummy-*it* referred to an antecedent in the previous sentence. Instead, *it is* should have formed the exceptional grammatical theme in (15) and *it should be questioned whether the claims made* the exceptional projecting and grammatical themes in (17). The coreferential *it* in (16) referred to *the phenomenon* in the previous sentence but was marked as a dummy-*it*. There, the *it* alone excluding the *was* should have been marked as the grammatical theme.

The cleft errors can be explained by means of the matching patterns defined for Spacy to find in the text. In Thematizer, specific dependency patterns together with lexical realization patterns were defined as a way to identify concrete structures. As outlined in Chapter 4.8, clefts can fall into non-projecting and projecting types (as in (15) and (17) above). Both clefts are introduced by a dummy-*it*, but the first is completed with a copula-adjective-infinitive structure. The second cleft, considered a projecting theme, may have an auxiliary modal but, in its base form, consists of a verb in passive voice followed by a projecting clause (*that*-clause or WH-adverbial). These two patterns were then defined as the established structures for Spacy to search for.

Complications arose, however, the more complex the predicate became. Sentences such as *It can be argued [however/regardless of implications/contrary to popular belief] that governmental reform is necessary* were not caught by the pattern matcher since an adverbial clause was inserted before the subordinating *that*-clause or adjective. Thematizer then assumed the dummy-*it* to be coreferential. Further, Thematizer assumed that two independent sentences were present if a conjunctive adverbial was inserted, such as *however* or *thus*. It then split the sentence and ignored the cleft structure entirely. Since the pre-defined Spacy pattern could not be expanded to include all permutations of a cleft structure, Thematizer failed to output the correct thematic parse.

Ultimately, failure to distinguish between projecting and non-projecting clefts constituted the dependency-based errors and resulting theme span misparses in this case. Similar to how t-unit misparses presented a significant error class, clefts also permeated every thematic parsing task due to coreferentiality and pattern-matching issues. While strides have been made in coreference resolution (see Stoyanov et al. 2009 and Stylianou & Vlahavas 2021), the errors that emerged in Thematizer's parses reflect the continued difficulty in tracing coreference chains. The complexity that underlies coreference resolution is made even more overt through the distinction necessary between projecting and non-projecting clefts, which possess remarkably similar syntactic structures. Often, contextual cues alone can account for whether

formulations such as *it was argued* are coreferential or pleonastic in nature. Despite the contextualized word embeddings in Spacy's language model, Thematizer could not leverage these in the resolution of coreference through the pre-defined pattern alone. As such, conflation of the dummy-*it* and coreferential *it* was a common occurrence, particularly in formal-register texts which made frequent use of projecting themes for objective language formulation.

Despite the diversity of error classes from the index identification task, their emergence was shown to ultimately stem from dependency parsing errors. As index identification was a syntactic task that revolved solely around the dissection of text into its smaller chunks – documents into paragraphs, paragraphs into sentences, sentences into themes and rhemes, themes into marked and grammatical themes – dependency parses were key to the analytical work that underpinned Thematizer's functionality and thematic analysis at this stage of the program.

6.1.1 Key Takeaways from Index Identification Parses

As Thematizer's parsing methodology for index identification was solely based on dependency tags and syntactic tests, the majority of the errors for this task stemming from dependency misparses suggests a deficiency in the programmatic approach. This then leads to the following key conclusion: While dependency parsing is a robust means of capturing the complexity of syntactic realization patterns, it alone cannot account for the identification of the clausal boundaries that constitute theme and rheme spans.

Clerical, grammatical and ambiguity errors proved to have a direct effect on the dependency parses, which ultimately hampered Thematizer's ability to accurately identify theme and rheme spans. The inclusion of additional testing parameters to reinforce the output from dependency tests could facilitate a reduction in thematic parsing errors, specifically through the inclusion of part-of-speech tagging.

The use of part-of-speech tags defined the core parsing approach in the work by Park and Lu (2015) and, to a degree, Domínguez et al. (2020). The former group of researchers were able to make use of the part-of-speech tags to generalize fourteen thematic patterns based on a rule-based system from Schwarz et al. (2008). These patterns became the basis for extracting theme and rheme spans with a resulting accuracy of F_1 = 93.0%. Domínguez et al., conversely, primarily employed dependency-based parsing for theme and rheme span extraction with an F_1 = 74.0%. Therefore, while Thematizer was able to exceed the accuracy of Domínguez et al. in index identification, the part-of-speech tagging approach by Park & Lu proved to be most accurate.

Automated part-of-speech parsing is not without error, as indicated by the F_1 = 93.0% score Park & Lu (2015) achieved; however, it appears to be less affected by clerical, grammatical and ambiguity noise that any text may have. Considering the case of semantic ambiguity raised in the previous section and reiterated here in Table 6-1, a dependency and part-of-speech parse would result in a biconditional testing parameter.

Text	*Intentionally*	*misleading*	*titles*	*vaporize*	*their*	*credibility.*
Thematicity	GRAMMATICAL THEME			RHEME		
Incorrect Dependency Parse	ADVMOD	AMOD	DOBJ	ROOT	POSS	DOBJ
Correct Dependency Parse	ADVMOD	AMOD	NSUBJ	ROOT	POSS	DOBJ
Part-of-Speech Parse	ADV : RB	ADJ : JJ	NOUN : NNS	VERB : VBP	PRON : PRP$	NOUN : NN

Table 6-1: Use of both a dependency and part-of-speech parse as a biconditional testing parameter for Thematizer to more accurately identify thematic spans with semantically and syntactically ambiguous propositional content.

Instead of relying on the output from the dependency parse alone, which failed to return an nsubj dependency in Thematizer's original parse, the part-of-speech parse could function as confirmation or refutation of the sentence constituents' assumed dependencies. The Penn Treebank tag (given after the colon in the part-of-speech parse in Table 6-1) for the grammatical subject titles is nns with the part-of-speech tag noun, indicating a plural noun. Since the verbal root vaporize was marked as third-person plural in the part-of-speech parse (verb : vbp), this could override titles as being misparsed as the direct object. This represents the first benefit to employing part-of-speech tagging as an additional testing parameter for index identification.

The second benefit is the generalization of part-of-speech patterns that can be used as queries when parsing sentences for their theme and rheme spans. Using the same sentence from Table 6-1, it reflects the standard, unmarked SVO structure. The noun phrase intentionally misleading titles could be reduced down to its head part-of-speech, noun, since intentionally misleading is dependent on the head titles. Additionally, for the isolation of the rheme, only the root is required since all constituents thereafter are automatically considered part of the rheme. Thus, the generalized pattern would simply be noun + verb with respective Penn Treebank tags, both of which would indicate the clausal boundaries for the theme and rheme. Since no sentence constituents syntactically independent from the noun appear in front of the titles, this pattern would suggest an unmarked theme. A similar approach could then be taken for the various realization patterns possible for sentences with marked themes. Dependency parses would then complement the part-of-speech parses returned for the sentence in question.

The importance of using a biconditional approach can also be seen in Thematizer's parsing methodology for clefts, which employed pattern-based matching only. Since the pre-defined pattern was overfitted, Thematizer was unable to account for deviations in the form of inserted adverbial phrases or parentheticals. Expanding Thematizer's pattern-based matching for clefts with part-of-speech and dependency parses could result in the more accurate identification of clefts with a wider range of realizational patterns. In fact, this assertion was already found to be true in Thematizer's marked theme classification, which employed both dependency parses and pattern-based matching and achieved the highest F_1 scores of all three thematic parsing tasks.

The final benefit to a two-pronged approach to parsing would be its ability to capture the diversity of realizational patterns across text types. Formal-register text types have a greater likelihood to adhere to conventional rules of grammaticality and syntactic structure, which dependency parses can capture more readily. Conversely, types with lesser formality can pose considerable difficulty for accurate dependency parses. For instance, the dependency parse produced for the text in Table 6-2 would have resulted in erroneous theme and rheme spans without consideration of the corresponding part-of-speech tagging.

Text	*Not*	*gonna*	*lie*	*dunno*	*how*	*come*
Thematicity	RHEME					
Incorrect Dependency Parse	NEG	ADVCL	XCOMP	INTJ	ADVMOD	ROOT
Correct Dependency Parse	NEG	ADVCL	XCOMP	ROOT	ADVMOD	CCOMP
Part-of-Speech Parse	PART : RB	VERB : VBG	VERB : VB	VERB : VB	SCONJ : WRB	VERB : VB

Table 6-2: Identification of thematic spans through dependency parses that have been reinforced with part-of-speech parses to account for sentence constructions in texts with a less formal register.

Parsing come as the root of the sentence would have caused Thematizer to assume dunno how to have been grammatical theme despite a lack of the nsubj dependency. In fact, the sentence in Table 6-2 has no grammatical theme due to elided grammatical subject congruent with *dunno*. Since the part-of-speech parse returned did not contain a NOUN, this information could have been compared against the dependency parse to prevent incorrect identification of a grammatical theme. Particularly in colloquial, informal texts with a greater frequency of contractions, ellipses and constituent completeness (in the sense of conventional grammaticality), both dependency and part-of-speech parses could accommodate the informal register and isolate clausal boundaries. As a reminder, Thematizer experienced particular difficulty with index identification in informal-register text types such as Reddit comments and lyrics, specifically. Such misparses could have then been mitigated if Thematizer's dependency parses had been reinforced with part-of-speech testing parameters.

In light of the findings above, Thematizer's accuracy for index identification ranging between 85.8% and 92.0% suggests a sound theoretical and programmatic foundation in the automatic extraction of theme and rheme spans in text. However, its failure to consistently reach the gold standard of 89.1% could be traced back to the myriad dependency misparses that emerged in both test datasets. For that reason, the operationalization of thematic theory in terms of theme and rheme identification was only partially achieved. Until Thematizer's accuracy exceeds that of the gold standard in its first parsing task, the reliability of Thematizer's index identification results should be considered with scrutiny. This could potentially be achieved by complementing Thematizer's current dependency-based parsing methodology with part-of-speech tagging for biconditional testing. A combinatorial approach, as shown to result in greater accuracy in previous work and Thematizer's marked theme classification, could both improve parsing accuracy and account for the myriad syntactic patterns that inform theme and rheme realization in discourse.

6.2 Key Findings and Error Classes: Marked Theme Classification

The next collection of error classes to discuss concerns marked theme classification and has been grouped according to underlying cause. The first major group of cascading misparses ties in with the errors discussed in Chapter 6.1 due to their original emergence in the index identification task. Because the underlying cause of cascading misparses was already elucidated in index identification, the effect they had on marked theme classification misparses will only be treated briefly. The second group of errors was caused by the misidentification or misuse of so-called right dependents, which were used to demarcate the terminating token of marked theme spans. This predominately affected circumstantial and hypotactic themes but also resulted in structural themes to be extraneously extracted as marked themes.

The defining error classes outlined below will be used to exemplify the underlying causes behind marked theme misparses. Chapter 6.2.1 briefly recaps the reasons for marked theme misparses originating from index identification errors. Chapter 6.2.2 then examines right

dependents as an underlying cause unique to marked theme classification. On the basis of these explanations and the findings from Chapter 5, conclusions are drawn on the operationalization of marked theme classification from a theoretical and programmatic perspective in Chapter 6.2.3. Final generalizations about Thematizer's ability to parse marked themes then conclude the treatment of automated marked theme classification.

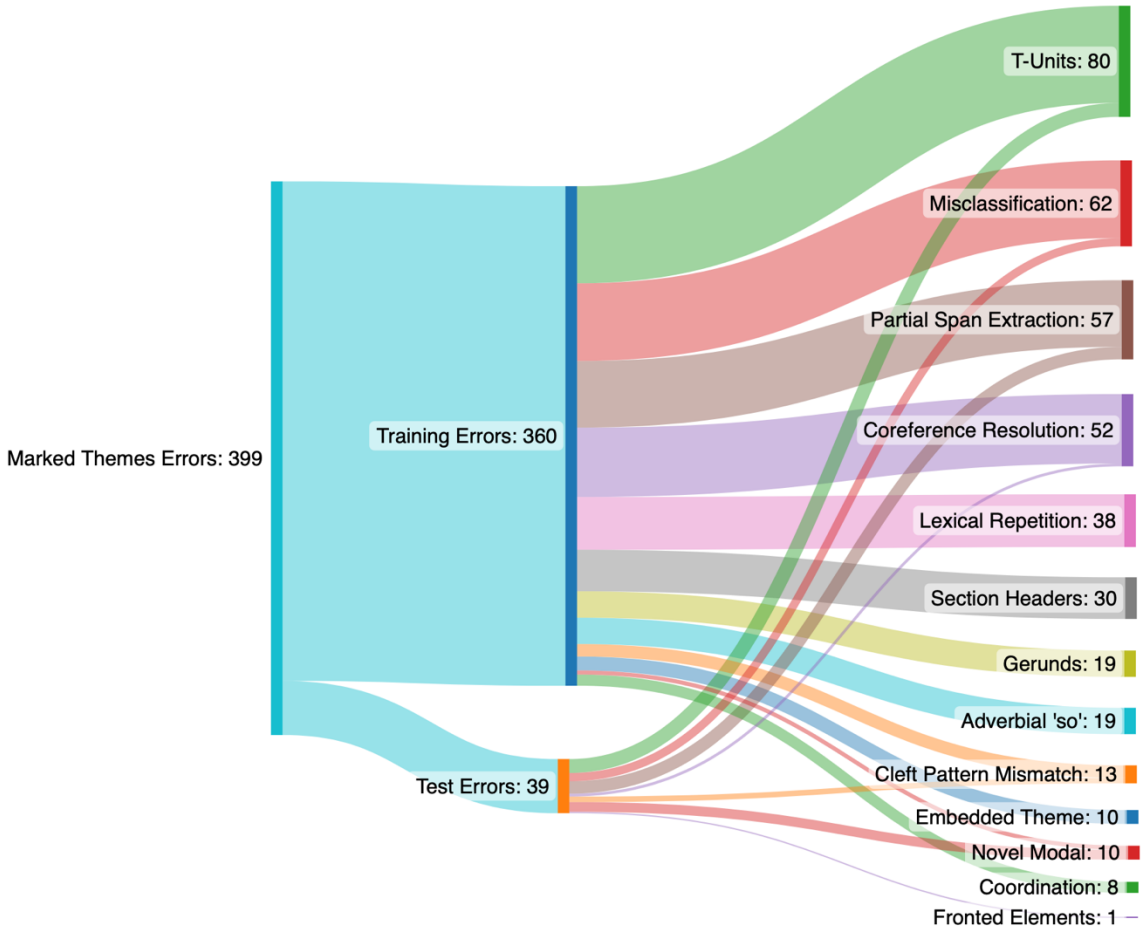


Figure 6-2: Summary of error classes in both training and test datasets for marked theme classification. Errors are given as absolute frequencies.

6.2.1 The Cascading Effect of Index Identification Misparses on Marked Theme Classification

Numerous error classes transferred across the two parsing tasks of index identification and marked theme classification, thereby affecting the subsequent parses that Thematizer had to attempt to resolve. As parsing errors from index identification became the input to process for marked theme classification, the resulting output was invariably erroneous as well. The specific key error classes that belong to cascading errors for marked theme classification from Figure 6-2 are: t-units, cleft pattern mismatches, coreference resolution, lexical repetition and fronted elements, section headers and gerund-induced misclassification.

Marked Theme Classification Misparses due to T-Units

T-unit misparses were, on the one hand, the result of Thematizer's failure to split two concomitant independent clauses joined by a semicolon, hyphen, colon, conjunctive adverbial or coordinating conjunction; on the other hand, falsely splitting a dependent clause from its independent, matrix clause constituted the group of false-positive misparses during index identification. Incorrectly parsed t-units then came to affect circumstantial, hypotactic and projecting themes specifically.

For circumstantial and hypotactic themes, a dependent clause separated from its independent, matrix clause resulted in fragments being introduced into the text to parse. Examples (1) and (2) show circumstantial and hypotactic misparses, respectively, as a result of falsely split t-units (indicated by the asterisk beside the t-unit split and thematic analysis).

(1)	T-UNIT 1	T-UNIT 2
*T-Unit Split	*They can more easily say hold on, just a minute, can I call you back?*	*Etc. without penalty or as much appeared disrespect.*
*Thematicity	GRAMMATICAL THEME (*they*) + RHEME (*can more...*)	CIRCUMSTANTIAL THEME + RHEME

In (1), the question mark embedded within the text caused both Spacy and Thematizer to assume two independent sentences. The clause after the question mark *etc. without penalty or as much appeared disrespect* then became a new t-unit and thereby newly introduced fragment. Subsequently, *without penalty or as much* was falsely identified as the circumstantial theme to the rheme *appeared disrespect*.

(2)	T-UNIT 1			T-UNIT 2	
*T-Unit Split	*After*	*Tallis*	*died in 1585.*	*Byrd*	*continued holding the patent.*
*Thematicity	HYPOTACTIC THEME	GRAMMATICAL THEME	RHEME	GRAMMATICAL THEME	RHEME

In (2), the pre-modifying dependent clause *after Tallis died in 1585* was split from its independent, matrix clause *Byrd continued holding the patent*. Here, Thematizer interpreted the adverbial *after* in the sense of *afterward* such that *Tallis died in 1585* was an independent, matrix clause. As such, Spacy split this clause from the actual independent, matrix clause *Byrd continued holding the patent*. The marked theme fragment was then parsed as a single sentence with its own thematic structure as indicated in the sentence parse analysis.

These two examples demonstrate the cumulative effect that misparses can have throughout subsequent parses. Since t-units were incorrectly split – or overlooked – during index identification, new thematic structures were introduced through their fragment structure. As incorrectly split t-units often started with circumstantial adjuncts, this additionally explains the prevalence of parsing errors for this marked theme type specifically. While not the primary contributor, t-unit misparses from the index identification task constituted 12.7% of the circumstantial misparses. For hypotactic themes, this number was 27.2%, which shows Thematizer's greater tendency to erroneously split sentences with hypotaxis than those with fronted prepositional phrases as in circumstantial themes. These splits are further substantiated through the occasional misidentification of the verbal root and congruent subject on account of the XCOMP and CCOMP dependencies from the fronted hypotactic clause.

T-unit parsing was also complicated through Thematizer's tendency to split dependent, projecting clauses from their independent, projecting matrix clause, particularly with clefts. This was the result of a cleft pattern mismatch, whereby a conjunctive adverbial or adjunct was inserted within the matrix clause. The insertion then prevented Thematizer from identifying a successful match for a projecting theme, as shown in (3). In such cases, the projecting theme was both overlooked and split as a separate t-unit on account of the conjunctive adverbial.

(3)	T-UNIT 1		T-UNIT 2		
*T-Unit Split	*Results*	*indicated.*	*However,*	*that these input variables*	*correlate with CIR defects.*
*Thematicity	GRAMMATICAL THEME	RHEME	STRUCTURAL THEME	GRAMMATICAL THEME	RHEME

Where conjunctive adverbials or adjuncts were inserted, Thematizer assumed that two independent sentences had been conjoined, which substantiated the t-unit split despite the subordinating conjunction *that*. Doing so then removed the original projecting theme entirely. Since t-unit parsing took place during text pre-processing before projecting theme extraction, Thematizer was wholly unaware of the projecting theme that had originally been present in the text. The reason for this t-unit misparse was an oversight in testing methodology: while the subordinating *that*-adverbial formed a test parameter for clause dependency in order to prevent such splits, it required the *that* to be sentence initial. In instances such as (3), however, the inserted adverbial became the sentence-initial constituent, such that Thematizer circumvented this testing requirement. As a grammatical subject and congruent verb were found within the projected theme (*these input variables correlate*), the parse was returned as an independent clause.

These findings reinforce the conclusion from the index identification task that Thematizer's parsing methodology for clause dependency testing was insufficient. While the presence of subordinating adverbials was included as test parameters for t-unit parsing, Thematizer occasionally circumvented this condition when non-subordinating adverbials started the dependent clause being parsed. T-unit misparses that emerged for this reason during index identification then manifested in circumstantial, hypotactic and projecting theme misparses.

Marked Theme Classification Misparses due to Cleft Pattern Mismatches

The second error class inherited from index identification to affect projecting themes was cleft pattern mismatches. The underlying cause here was due to the MARK dependency used as a test parameter for potential cases of projecting themes. This then caused related misparses stemming from the *that*-clause approximants *so…that* and *such that*. Each of these misparses will be detailed in the following examples, whose sentence constituents in bold form the misparsed projecting theme.

(4) *** REM's Murmur was so out of the moment that** it simply had to resonate with anyone looking for an alternative.

In (4), Thematizer identified the subordinating *that* on account of its MARK dependency, which was used to differentiate it from the coreferential or relative pronoun *that*. If a subordinating *that* with MARK dependency was identified, then Thematizer assumed that the sentence contained a projecting theme. The relevant sentence constituents were then extracted and subsequently parsed as a projecting theme, albeit erroneously. In cases such as (4), Thematizer should have marked *REM's Murmur* alone as the grammatical theme instead of belonging to a projecting theme.

The same case is found in (5), where the non-projecting approximant *such that* should have precluded projection parsing. Instead of *oviposition* alone being parsed as the grammatical theme, Thematizer marked everything up to and including *such that* as the projecting theme due to the MARK dependency.

(5) * **Oviposition has been observed on a wide range of native and introduced plant species and can weaken the branches of young orchard trees such that** they cannot sustain the load of their fruit.

The lack of projection in this example can further be explained by the lack of a projecting verb together with a *that*-adverbial. The final verb *weaken*, on which the approximate *such that* is syntactically dependent, does not fall under the category of verbal, relational or behavioral processes in verbal clauses (Halliday & Matthiessen 2014: 134). The reason for this is the interpersonal function that projecting clauses inform. Projected clauses can express viewpoint explicitly (e.g., *I believe that…)* or implicitly (e.g., *It could be argued that…*). While the explicit realization is modal in nature, the implicit formulation is objectified and presented as factual. As such, the verb used in the projecting class must fall under an interpersonal or experiential class function.

In the case of (5), *[oviposition] can weaken the branches* is neither an implicit nor an explicit interpersonal formulation on account of *weaken*: the expressions **It can be weakened that…* or **They weakened the branches that…* would be infelicitous in a projecting sense: only the second expression would be acceptable through the referential use of *that* as a relative pronoun. Furthermore, *such that* and the related *in that* serve to introduce elaboration or clarification of the information presented beforehand. The theme of sentences with such approximants is fully realized and discursively GIVEN (i.e., not introduced through a cleft or dummy-*it* structure), as evident in (5). As such, the approximants *such that* and *so that* should have been excluded from the test condition of a MARK dependency when identifying potential projecting themes.

These errors indicate a fallacy in the reliance on the MARK dependency as a test parameter for the identification of projecting themes. While the use of MARK was able to account for actual projections, failure to delineate the various forms of the subordinating *that*-adverbial resulted in an overgeneralization of projecting themes. Expanding Thematizer's test parameters for projection to ignore *that*-adverbial approximants would prevent the parser from identifying them as projecting themes.

Projecting themes were further compromised due to the pre-defined search pattern having failed to identify projecting clefts with inserted elements. In such cases, the *it* + *copula* + *participle* + *that* pattern was interrupted by an adverbial insertion. As such, Thematizer failed to mark the construction as a projecting theme, which resulted in the dummy-*it* becoming the grammatical theme. As such, the projecting theme that should have identified was overlooked.

Incorrect Parse	(6) *It*	*is, unless stated otherwise, assumed that out-of-spec output variables are indications of rubber failures.*
Incorrect Thematicity	*GRAMMATICAL THEME	*RHEME

Corrected Parse	(6) *It is, unless stated otherwise, assumed that*	*out-of-spec output variables*	*are indications of rubber failures.*
Corrected Thematicity	PROJECTING THEME	GRAMMATICAL THEME	RHEME

As long as the cleft pattern was unable to account for divergent cleft patterns, as in (6), then Thematizer was unable to parse it as a projecting theme. Expanding the cleft's possible realizational patterns as long as the basic *it* + *copula* + *participle* + *that* is fulfilled would sufficiently generalize the pattern and thereby increase its coverage in the parses. This approach would also fall in line with previous pattern-based approaches, whereby basic sentence constructions are reduced to their part-of-speech tags for identification, as outlined in Chapter 6.1.1 on index identification.

Marked Theme Classification Misparses due to Lexical Repetition & Fronted Elements

Next, lexical repetition and fronted elements were often assumed to be circumstantial themes or modal themes during index identification. This was due to commas affecting the resulting dependency parse such that dependency of the repeated lexeme or fronted element was linked with the verbal root of the matrix clause. Lexical repetition errors most commonly affected circumstantial themes in the text type of lyrics specifically. In (7), the repeated phrase *Ice Age* caused Thematizer to assume the fronted noun phrases to be circumstantial and of type LOCATIVE.

(7) **Ice Age**, Ice Age's coming.

Similar to the example (6) in Chapter 6.1 with *bad, bad medicine*, the dependency for the repeated lexeme or lexical phrase was associated with the verbal root *coming*. Thematizer then qualified *Ice Age* as a marked theme since it appeared before the second instantiation of *Ice Age* as the subject. While elements appearing front of the grammatical subject and offset by a comma should indeed be parsed as dependent on the matrix' verbal root, that syntactic parse should only hold true when the fronted elements are independent of the grammatical subject. Since *Ice Age* itself is the grammatical subject and therefore dependent on its second instantiation, its repetition should have prevented Spacy and Thematizer from marking it as a separate dependent clause. This error class's frequency of 25.8% indicates how pervasive it is in dependency parses, which a pattern-based approached could obviate: a repetition of the noun phrases immediately followed by the verbal root would result in an unmarked grammatical theme + rheme structure. That would then prevent the first repeated or fronted element from being passed on to marked theme classification.

Fronted elements in the form of vocatives and interjections resulted in modal themes being entirely overlooked or misidentified due to their novel use. Novel, here, means that the modal theme was not saved in the pre-defined look-up tables for modal themes. When Thematizer attempted to parse them, then, no match was found and they erroneously defaulted to the grammatical theme. For example, in *Ah, ty for the info*, neither *ah* nor *ty* was marked as modal themes due to their absence in the look-up tables. Novel modals such as *honey* in the vocative *Honey, I'll surrender* were overlooked for the same reason. Adding these modal themes to the look-up tables resulted in their correct parse, which indicates Thematizer's fundamental ability to account for them. As is the case with all look-up tables, continuing to populate them with thematic realization patterns forms the basis of Thematizer's future development.

Marked Theme Classification Misparses due to Header Misparses

The penultimate error to have stemmed from index identification was the misparse of headers (cf. Chapter 6.1). In section headers, often without a grammatical subject or finite verb, syntactic and semantic ambiguity resulted in clauses being mistaken for independent clauses

with an assumed grammatical subject and verbal root. An additional common misparse was the assignment of numerals to circumstantial themes since they appeared sentence-initially.

(8)	*3.*	*Round up your purchases.*
Token Index Span	[0–1]	[2–6]
Incorrect Parse	*CIRCUMSTANTIAL THEME	*RHEME
Corrected Parse	RHEME	

In (8), the sentence-initial *3.* was denoted as the marked theme despite the lack of a grammatical theme. This occurred because of the test parameter for marked themes, whose ending index had to be greater than the starting index of the sentence or clause. This then allowed the entirety of the marked theme span to be extracted and passed on for subsequent parsing. The unintended side effect was the inclusion of numerals to be parsed as circumstantials since their ending index (1 in the example above) was greater than the starting index of the sentence (0 in (8) above). This error then specifically stems from an oversight in the test condition, which overgeneralized how sentence-initial elements should be processed, or, in the case of (8), ignored. Correction of this testing parameter would then ameliorate the erroneous parsing of enumerations as circumstantial themes.

Marked Theme Classification Misparses due to Gerunds or Participle Phrases

The final error class to have been inherited from index identification was a particular form of misclassification that affected hypotactic themes alone. This occurred when nominalized verbal phrases, i.e., gerunds or participle phrases, functioned as the grammatical subject but were identified as non-finite relative clauses. This was partially caused by commas, correct or not, and Thematizer's occasional inability to identify gerunds as grammatical subjects.

Incorrect Parse	(9) *Alternatively,*	*induced maternal hyperglycemia,*	*by continuous glucose infusion, causes fetal hyperglycemia [...]*
Incorrect Thematicity	STRUCTURAL THEME	* HYPOTACTIC THEME	*RHEME

Corrected Parse	(9) *Alternatively,*	*induced maternal hyperglycemia, by continuous glucose infusion,*	*causes fetal hyperglycemia [...]*
Corrected Thematicity	STRUCTURAL THEME	GRAMMATICAL THEME	RHEME

In (9), the participle phrase *induced maternal hyperglycemia* is followed by the parenthetical *by continuous glucose infusion* offset by commas. Here, it was the parenthetical and the participle *induced* that prevented correct identification of the subject. The participle was incorrectly parsed as ADVCL, which "modifies a verb or predicate" and not the subsequent nominal phrase (Nivre 2022). This was due to *maternal hyperglycemia* being assigned the dependency of direct object (DOBJ) as a nominal complement of the transitive *induced*. That implies that *maternal hyperglycemia* was not ascribed the semantic role of patient to indicate it underwent the action of *induce*. Since *maternal hyperglycemia* was erroneously afforded direct object status instead and the subsequent parenthetical prepositional phrase could not be the grammatical subject, the entirety of the phrase was denoted as a marked theme. Yet again, the use of pattern-based matching as a failsafe or secondary test to the dependency parse would have been one way of confirming or refuting the output produced. Doing so would have

overridden the false direct object parse to equate it with the congruent verb *causes* in the example above.
The discussion of the cascading effects in the present section highlighted how misparses during index identification then manifested in marked theme classification misparses. While most errors were attributable to t-units, cleft pattern mismatches and lexical repetition errors, less frequent were then those pertaining to fronted elements, section headers and gerund-induced misclassification. Each of the cascading errors affected one or more of the marked themes but circumstantial, hypotactic and projecting themes were impacted most on account of the underlying causes. Most of the errors were shown to be a result of faulty dependency parsing as first outlined in Chapter 6.1. Yet, test conditions for assumed marked themes in section headers and subordinating *that*-adverbials not appearing sentence initially in projecting themes proved an oversight in the programmatic approach. Just as misparses from index identification resulted in misparses in marked theme classification, resolving these issues in the first parsing task would mitigate or, at best, negate their emergence in the second.

6.2.2 Marked Theme Misparses from Right Dependents

The present section turns to the major underlying cause behind marked theme misparses unique to the marked theme classification parse. This case was so-called right edges or dependents, whose indices were used to determine the end of marked theme clauses. Depending on the kind of marked theme, however, the right dependent of the adjunct's syntactic head was required for successful end-of-span identification. Errors stemming from misidentification of right dependents were misclassification, partial theme extraction, coordination and embedded themes. While most errors affected circumstantial and hypotactic themes, structural themes were also found to be occasionally affected by right-dependent errors.

To help explain what these indices are and how they were used in Thematizer's parsing methodology, the following two visualized dependences have been provided. These will aid in understanding how misparses arose during marked theme classification.

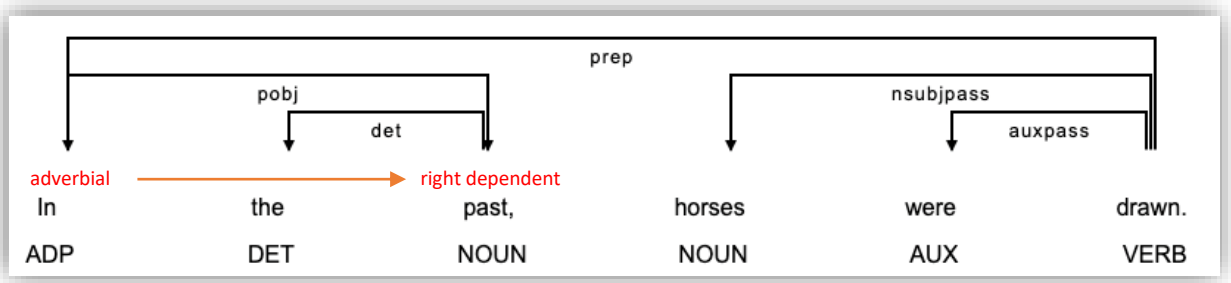


Figure 6-3: Visualization of the dependency parse with the circumstantial theme in the past *to illustrate right dependents' use in determining the end of a marked theme span. Here, the right dependent of the circumstantial adverbial* in *is* past.

In the first example from Figure 6-3, the circumstantial theme *in the past* serves as the introductory phrase dependent on the matrix clause *horses were drawn*. As is the case with nearly all circumstantials, the adverbial *in* is a descendent of the matrix' finite verb *drawn* as indicated by the left-pointing arrow drawn between the two constituents and dependency PREP. Then, the circumstantial theme possesses its own syntactic tree, whereby the object of the preposition *past* forms the right dependent, as shown in red and indicated by the right-facing arrow to *past*. Hence, by identifying the index of the right dependent, Thematizer was able to isolate the circumstantial theme in its entirety.

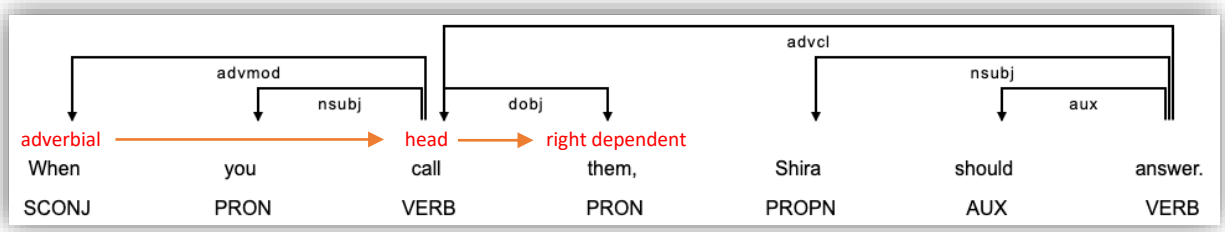


Figure 6-4: Visualization of the dependency parse with the hypotactic theme when you call them *that illustrates the necessity of the right dependent of the adjunct's head to mark the end of the marked theme span. Here, the head of the adjunct* when *is* call*, whose right dependent is* them.

However, as soon as a subordinate clause functions as a marked theme, such as through the hypotactic theme *when you call them* in Figure 6-4, the right dependent alone does not suffice. The reason for this is because the right dependent of the hypotactic adverbial *when* is *when* itself: this is indicated by the left-facing arrow above *when*, whose origin is the finite verb *call* within the subordinate clause. Therefore, *when* is a descendent of *call*; the converse relationship, then, is that *call* is the head of *when*. By accessing the adverbial's head, which is always the finite verb from within the subordinate clause, the token that terminates the marked theme span can be identified through its right dependent. For hypotactic themes in particular, it is the right dependent of the adjunct's syntactic head that is the critical syntactic node. At this point, it should also be noted that any post-modification to the head of a phrase, e.g., through a relative clause, appositive or parenthetical, can also only be accessed through the right dependent of a head. For example, if the circumstantial theme in Figure 6-3 were *In the past, which is not now…*, Thematizer would have to make use of the right dependent of the syntactic head to extract the entirety of the marked theme.

Marked Theme Misclassification due to Right Dependent Misparses

With this distinction in mind, a discussion of how and why Thematizer failed to account for right dependents in marked theme classification can ensue. The first effect of right dependent misparses was the misclassification of circumstantial themes. In (1), the sentence begins with a temporal circumstantial theme that has been complemented by an additional subordinated temporal phrase.

(1) **On June 21, 2017, three days after "Coal" aired,** Marshall County Coal Company and other companies chaired by Murray filed a strategic lawsuit.

Since post-modification of the circumstantial theme *on June 21, 2017* occurred, Thematizer made use of the right dependent of the head, i.e., *aired* not *2017*. Therefore, the circumstantial span was successfully extracted in its entirety. The resulting semantic classification was LOCATIVE but should have been TEMPORAL. The reason for this was how Thematizer categorizes marked themes into their semantic subclasses.

Here, a combination of syntactic tests and pattern-based matching were used to determine class membership and to differentiate between multiple semantic class options where appropriate. If the dependency parse of the right dependent was returned as a numeral for circumstantial themes, then Thematizer qualified the marked theme of class TEMPORALITY. Otherwise, if the marked theme was an established, non-modifiable phrase, such as *in addition* or *conversely*, pattern matching automatically ascribed the corresponding semantic class without the need for syntactic parsing.

Where problems arose, then, was with adjuncts whose semantic subclass had to be disambiguated on account of multi-class membership. For instance, *on* can either be LOCATIVE as in *on the desk* or MATTER through *on the topic of lexicality*; similarly, *with* can either be ACCOMPANIMENT as in *with me* or MANNER as in *with this tool*. Multiclass membership thereby increased the likelihood that Thematizer would categorize the circumstantial theme incorrectly since it had to rely on the returned dependency parses.

Returning to example (1), the right dependent of the head *aired* had a dependency of adverbial clause (ADVCL), indicative of its appositive structure subordinate to the circumstantial prepositional phrase. Since ADVCL was returned and not a numeral, Thematizer assumed that the entirety of the phrase was LOCATIVE, which was the default parse if no other dependency parses matched the test conditions for semantic class.

Thematizer's dependency on syntactic tests with the right dependent caused further misclassification errors when the returned dependency was inadequate for disambiguation. Whereas the appositive in (1) caused the wrong dependency to be used for classification, it was the semantics of the head of the preposition in (2) that determined the corresponding semantic class, not the syntactic dependency.

(2) **In a vote characterized by intimidation**, the 19 May 1916 referendum on whether to change the city name decided "yes" by a slim margin.

In other words, the semantic class of the circumstantial theme should have been derived from the semantics after *in a vote*, which would have equated to MANNER (i.e., by means or through the vote) instead of LOCATIVE as Thematizer returned. The LOCATIVE misparse was then on account of the dependency parse object of the preposition (DOBJ) and part-of-speech tag NOUN. As semantic tests were not included in marked theme classification, Thematizer had to rely on dependency parses alone for semantic classification. Only 62 of the 1046 circumstantial themes experienced misclassification due to unresolved disambiguation, which indicates the strength in Thematizer's current programmatic approach and the relative infrequency of required classification on the basis of a lexeme's semantic contribution alone. That being said, semantic classification of marked themes belonging to multiple semantic classes requires further testing methodologies in future developments of Thematizer.

Partial Extraction of Marked Themes due to Right Dependent Misparses

Right dependent misparses were further complicated by a particular subset of circumstantial themes belonging to the class of temporal adjuncts as noun phrases. As a visualization of the dependency parse for such cases illuminates right dependencies, the example in Figure 6-5 will serve as the basis for discussion of noun-phrase circumstantials.

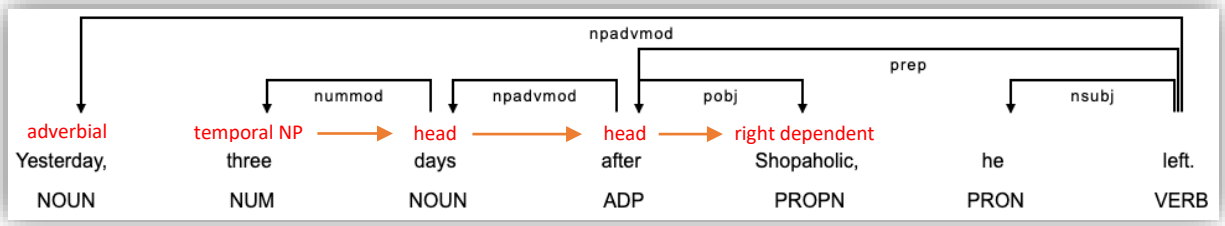


Figure 6-5: The dependency parse for post-modified temporal noun phrases, which has a doubly nested right dependency. Here, the head of the circumstantial adjunct initializer three *is* days, *whose head is* after. *The right dependent of* after *is then* Shopaholic *the end of the circumstantial theme.*

Firstly, the sentence in Figure 6-5 contains the two circumstantial themes *yesterday* and *three days after Shopaholic*. Single-word temporal circumstantials are their own right dependent since they possess no syntactic children but instead are descendants of the matrix' verbal root, here the sentence-terminating *left*. The second circumstantial, however, is more complicated in that the individual constituents *three days after* form a single temporal noun phrase but are descendants of one another. When compared to the standard circumstantial theme dependency in Figure 6-3, the red right-pointing arrows indicate ascendency, which allows the right dependent to be used directly for theme span extraction. With temporal noun phrases for circumstantial themes, therefore, multiple head nodes must be traversed until the right dependent can be identified as the end of the phrase. In Figure 6-5, the head of *three* is first isolated (*days*), followed by its head (*after*), at which point the right dependent *Shopaholic* can be extracted as the clause-terminating token.

It was due to the inherently nested structure of temporal noun phrases as circumstantials that Thematizer was unable to extract the marked theme span in its entirety since only the first right dependent of the head, i.e., *days* in Figure 6-5 was extracted. While that then resulted in the correct semantic classification of *three days* as TEMPORAL, the failure to extract the whole marked theme led to some constituents remaining overlooked.

The same case was evident in certain hypotactic themes, whereby the final right dependent of the adjunct's head was not accounted for. In (3), a compound infinitive phrase formed the hypotactic theme, whose clause-terminating right dependent was *Jerusalem*. In this case, Thematizer only extracted the first infinitive phrase *to consolidate his hold on the city* due to the nested and coordinated hypotactic themes that followed.

(3) ***To consolidate his hold on the city,** monitor events on the Temple Mount and safeguard the Hellenized faction in Jerusalem, Antiochus stationed a Seleucid garrison in the city.

Issues for hypotactic themes such as these were further complicated through the failure of an end-of-clause condition. In order to ensure that the right dependent was actually found, Thematizer tested for a clause-terminating comma, such as the comma between *Jerusalem* and *Antiochus* in (3). If found, then the right dependent of the head was confirmed as correct. Occasionally, however, commas were not correctly identified due to text encoding issues. While the entirety of the text was encoded as UTF-8 to ensure uniformity, punctuation in particular would sometimes fail to be encoded correctly. This resulted in punctuation conditions as a test parameter failing since the input comma did not correspond to the comma encoded as UTF-8 in Python's collection of pre-defined string punctuation.

Marked Theme Misparses due to Embedded Themes and Coordination

The culmination of right dependent complexity was found in the hypotactic themes *so as to* and *so that*, the former of which has been visualized in Figure 6-6. While belonging to the class of infinitives, their dependency indicates a deviation from other related forms of infinitive phrases, such as *in order to, in order for* and the non-finite *to*.

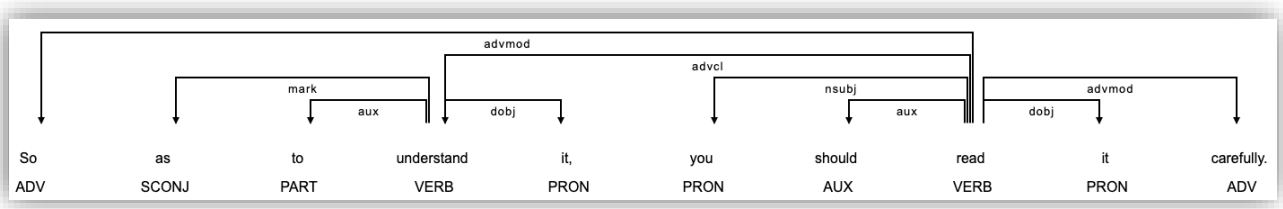


Figure 6-6: Dependency visualization of the hypotactic theme introduced by so as to*, which requires a compound right dependent parse and right dependent of the adjunct's head for correct extraction.*

First of all, instead of being parsed as a compound dependency to produce the established phrase *so as to*, the dependency parse marked *so* as an adverbial that is dependent on the matrix' verbal root *read*. For that reason alone, access to *as* or *to* could not be achieved through a right dependent or right dependent of the head. As such, Thematizer consistently and erroneously extracted *so* first as the structural theme and then parsed the residual *as to understand it* as a hypotactic theme. The parse of hypotactic themes with *so that* resulted in even more infelicitous parses once Thematizer extracted the sentence-initial *so*. Since the residual hypotactic, e.g., *that you understand it*, did not correspond to syntactic patterns for any marked theme type, Thematizer defaulted to marking the theme as a grammatical theme. This had the added drawback of introducing new structural themes to the text output that syntactically belonged to parent nodes as established phrases.

This tendency to introduce new structural themes on the basis of right dependent parsing was also found where coordination occurred within marked themes. In such instances, the right dependent alone failed to correctly demarcate the end of the marked theme span, such that the token occurring before a coordinating conjunction was identified as the right dependent. This is shown in (4), which contains coordination within a hypotactic infinitive phrase.

(4) *** To analyze the difference** and methods of resolution it is fundamental to mention that 5% of children in Europe don't have a suitable place to do homework and 6,9% (*sic*.) have no access to the Internet.

Here, the right dependent of the hypotactic-initiating head *to* is *resolution*. However, Thematizer incorrectly identified *difference* as the right dependent, which resulted in *to analyze the difference* alone being extracted and partially identified as a hypotactic theme. In the subsequent recursive step, *and methods of resolution* was returned for processing. Since this phrase began with *and*, Thematizer processed it as a structural theme regardless of its realization within the coordinated infinitive phrase. In the final recursive step, *methods of resolution* was returned, which did not correspond to any marked theme structure. As such, this final noun phrase was unable to be parsed and was returned as the default grammatical theme.

Instead, coordination embedded within dependent clauses – hypotactic or otherwise – should have been ignored on account of the right dependent. It should be noted that overgeneralization of coordination occurred primarily within hypotactic themes followed by a matrix clause without a grammatical subject. This was due to the assumption that the matrix clause was dependent on, not independent from, the hypotactic clause. The upper dependency parse in Figure 6-7 illustrates a compound hypotactic theme with a matrix clause lacking a grammatical subject. The lower parse is the same sentence but with the grammatical subject *you* included in the matrix clause.

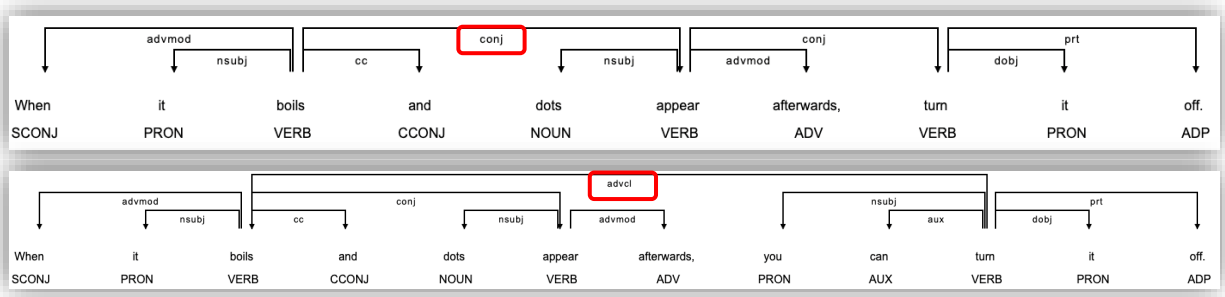


Figure 6-7: Overgeneralization of coordination within hypotaxis, which causes dependency parses to assume continued coordination within the matrix clause. This then causes the matrix clause to become dependent on the hypotactic clause, not independent from it, as indicated by the red dependency parses CONJ *and* ADVCL.

The key difference between both dependency parses is the concatenation of the coordinating dependencies through CONJ in the upper parse compared to the adverbial clause dependency ADVCL in the lower parse as marked in red. It is the correct ADVCL dependency that indicates the matrix clause's independence from the hypotactic clause. The concatenated CONJ dependencies in the upper parse, however, illustrate how the parser assumed *turn it off* to be a continuation of the hypotactic clause. In other words, the dependency parse assumed that three coordinated dependent clauses (*it boils*, *dots appear* and *turn it off*) were present. For that reason, the right dependent of the hypotactic adjunct *when* was returned as the sentence-terminating token *off*. If the matrix clause had been parsed as independent from the hypotactic clause, as in the lower parse, then it would have returned the correct right dependent of the head as being *afterwards*.

While this is a very specific case and only occurred in 0.8% of the errors, it illustrates how critical correct dependency parses are for Thematizer's extraction of marked themes via right dependents. It also indicates how minor changes to sentence structure – the mere absence of a grammatical subject in the matrix clause – impact the resulting dependency parse. Right dependents are a robust syntactic means for capturing the near infinitude of syntactic realization patterns in marked themes as indicated by their high parsing accuracy. However, this example with coordination sheds light on the periphery cases that can remain unaccounted for when dependency parses are faulty.

All in all, right dependents forming a primary cause for misparses of hypotactic and circumstantial themes are a reflection of the inherent complexity these marked themes can possess. Through their own grammatical subjects, finite verbs, potential adjuncts and complements, hypotactic marked themes have the greatest degree of syntactic complexity. Circumstantials, which are most often realized as prepositional phrases or temporal noun phrases, can also achieve varied and complex realizational patterns through additional subordination from relative clauses and complements. If dependencies within these marked themes are incorrectly analyzed, then the use of right dependents as an end-of-clause marker for thematic parses invariably suffer. Ultimately, marked theme misparses in Thematizer's analysis of circumstantial, hypotactic and (overgeneralized) structural themes were thus commonly induced through partial theme extraction and misclassification of coordinated structures and embedded themes.

6.2.3 Key Takeaways from Marked Theme Classification

Errors that affected Thematizer's parsing accuracy of marked theme classification were shown to originate from two key groups: firstly, cascading errors from the index identification task;

and secondly, right dependents causing partial extraction or misclassification of circumstantial, hypotactic and occasionally structural themes. Deficiencies in Thematizer's parsing methodology came to the fore when demarcating the spans of marked themes, grammatical themes and rhemes during index identification. With its second thematic parse, Thematizer had to classify marked themes on the basis of potentially erroneously identified marked theme spans. It was these errors that pervaded nearly 70.0% or more of marked theme errors. The remaining were caused by the misidentification of right dependents, whose use, while theoretically straightforward, was impacted to varying degrees depending on the syntactic complexity of the marked theme.

While the underlying causes outlined in Chapter 6.2.1 and 6.2.2 revealed missteps in certain test parameters and conditions, the overall accuracy Thematizer achieved in marked theme classification (F_1 = 94.9% for training data; F_1 = 93.4% for test data) indicates its successful operationalization via automated and computational means. As concluded in Chapter 6.1.1, Thematizer's employment of both dependency and pattern-based tests in marked theme classification proved pivotal in the resulting parses and corresponding accuracy rates. The higher accuracy for marked theme classification compared to index identification strengthens this finding as index identification relied on dependency parsing only.

This finding is similar to the approach to extracting thematic constituents in Park & Lu (2015), who achieved an F_1 score of 93.0% through dependency and pattern-based parsing. Further, Puşcaşu et al. (2006) were able to automate the extraction of temporal connectives using a multivariate approach: machine learning methods were trained on cue phrases, i.e., the adverbial that initiated or defined the temporal connective, such as *when* or *often*, which were annotated with their semantic class. Then, dependency parses were used in the training model for the identification of temporal connectives in text. Their approach was closest to that of the present work, although machine learning was not employed in the development of Thematizer. Ultimately, Puşcaşu et al. (2006) were able to achieve an accuracy of 89.2% with their model. The results from their work and from Thematizer seem to suggest that high parsing accuracies can be achieved through a combinatorial testing methodology: dependency parses form the foundation of the identification and extraction of adverbial clauses as marked themes; then, through pattern-based matching, classification of the marked themes' semantic class can be determined.

That being said, dependency-based parsing alone appears to be the primary approach that most previous work has employed and that has achieved reliable accuracy rates. Chen et al. (2021) focused on all forms of subordinate clauses and used universal dependencies to isolate their use in text with an accuracy of F_1 = 93.8%. Similarly, Chen & Maning (2014) made use of so-called transition-based dependency parsing with neural networks. This allowed the prediction of part-of-speech tags, the syntactic head of the adverbial and its dependency parse, the verbal root as the parent node of the syntactic tree and the position of the word in the parsing stack (Chen & Maning 2014: 741). Through training and test validation, they achieved accuracies between datasets of 90.7% and 92.2%.

Regardless of underlying parsing methodology, all accuracy results, including that of Thematizer, revolved around the 93.0% F_1 mark. For Thematizer specifically, the minimum 93.4% accuracy demonstrates the reliability of the analytical output returned to the user. This is particularly important when gaining impressions of or drawing conclusions on marked theme use in written text. As Thematizer was able to reliably identify, extract and classify marked themes into their semantic subclass, greater detail into their frequency, diversity in use and role in text register was afforded to the accompanying thematic results. This achievement was made

because of, or perhaps despite, the added complexity that parsing marked themes into different syntactic and semantic classes required.

Thematizer's marked theme parsing functionality additionally represents an improvement over previous approaches to automated thematic analysis. While Park & Lu (2015) identified the metafunctions that corresponded to the thematic elements – experiential, textual or interpersonal – no further classification was offered. This revealed how metafunctions were realized thematically in text; however, their analysis remained largely abstract as no further delineation of the metafunctions occurred in terms of their syntactic or semantic contribution to the text. Thematizer, conversely, was not only able to identify marked (and unmarked) themes, but also qualify their contribution to text structure and contextualization of the discourse message in each sentence via functional categories (marked theme types) and semantic classes. Of all three parsing tasks, marked theme classification therefore represents the key candidate for Thematizer's successful operationalization of thematic structure. It further indicates how Thematizer was able to overcome deficiencies in the analytical output of previous work, which formed the motivation behind the first research question in the present work.

The final point to be addressed concerning marked theme classification is their frequency distribution with respect to register (cf. Chapter 5.3). In the present work, it was found that more complex marked themes (circumstantial, projecting, and, to a degree, hypotactic themes) were more frequent in text types of a formal register. Simultaneously, more simplistic marked themes, such as those reflected in structural and modal themes, were more frequent in text types of a less formal register. Therefore, marked theme complexity appears to be a reflection of text type formality.

This finding is supported by Hasselgård (2010) who investigated the frequency, position and distribution of adjuncts with respect to thematization and text type. While adjuncts related to space and time were equally frequent across all text types, the formal-register text types investigated (news and academic texts) exhibited the greatest use of a wider range of adjuncts for discourse-contextualization purposes (Hasselgård 2010: 269). The higher frequency of marked themes with greater syntactic complexity in formal texts can be seen as a reflection of the subject matter's complexity, particularly as sentence-initial adjuncts. The purpose of fronting adjuncts and thereby realizing marked themes is the framing and situating of the discourse message to follow in the sentence. Established discourse topics are employed in marked themes as an initial foundation for the reader to base the rhematic development of the message on. Circumstantial, projecting and hypotactic themes then allow these discourse topics to be captured and continuously re-contextualized through the unfolding of a text. As more formal text types weave a cornucopia of discourse topics throughout a text's unfolding, marked themes that allow for greater syntactic realizational patterns facilitate a reinforced scaffolding for discourse development.

The same holds for less formal text types, which more readily rely on structural and interpersonal connectors emblematic of conversation and spontaneous speech (Hasselgård 2010: 278). As shown in his seminal work, Biber found that coordination was a key characteristic of spoken speech as a means to concatenate the expression of thought in a linear and simplistic manner (1995: 106–107). Further, the inherent interpersonal nature of conversation enjoys greater use of adjuncts that indicate modality, perspective, emotive expression and stance (Biber 1995: 47). In the present work, both coordination and interpersonal modality was found to be the case in the less formal text types, whose coordinating and conjunctive connectors as well as modal adjuncts constituted 58.1% of the marked themes (see Figure 5-7 in Chapter 5.3). Therefore, while the analyzed texts were of the written mode, the less formal text types reflected

a higher number of texture characteristics commonly associated with spoken speech. Just as objective, factual language is employed in more formal text types to establish distance between the author and the reader, less formal text types leverage spoken-speech-like characteristics to increase the degree of interaction with the reader and their accessibility to the text.

The output that Thematizer produces for a text's marked theme distribution can therefore provide insight into where a user's text falls on the cline of register vis-à-vis marked theme use. This can be of particular benefit to non-native writers who may employ a narrower range of marked themes to develop their texts or at a frequency atypical of the text type they are producing. For writers of any background, the frequency distribution of their marked theme usage can reveal particular idiosyncratic patterns in their writing they may wish to reduce (e.g., the sole use of *in addition* as an elaborative structural theme) or develop further. Comparative analyses between the user's own text and those of other text types, registers or genres could further raise an author's awareness of marked theme usage from an intertextual perspective. Due to the accurate output that Thematizer produces, users can draw reliable and more generalizable conclusions about marked themes in their writing and across text types. They can then use this information as automated feedback for continued improvement to their composition, their writing style and the development of their texts.

6.3 Key Findings and Error Classes: Thematic Progression

The underlying causes behind the key error classes for Thematizer's third and final thematic parse, thematic progression classification, constitute the discussion of the present section.

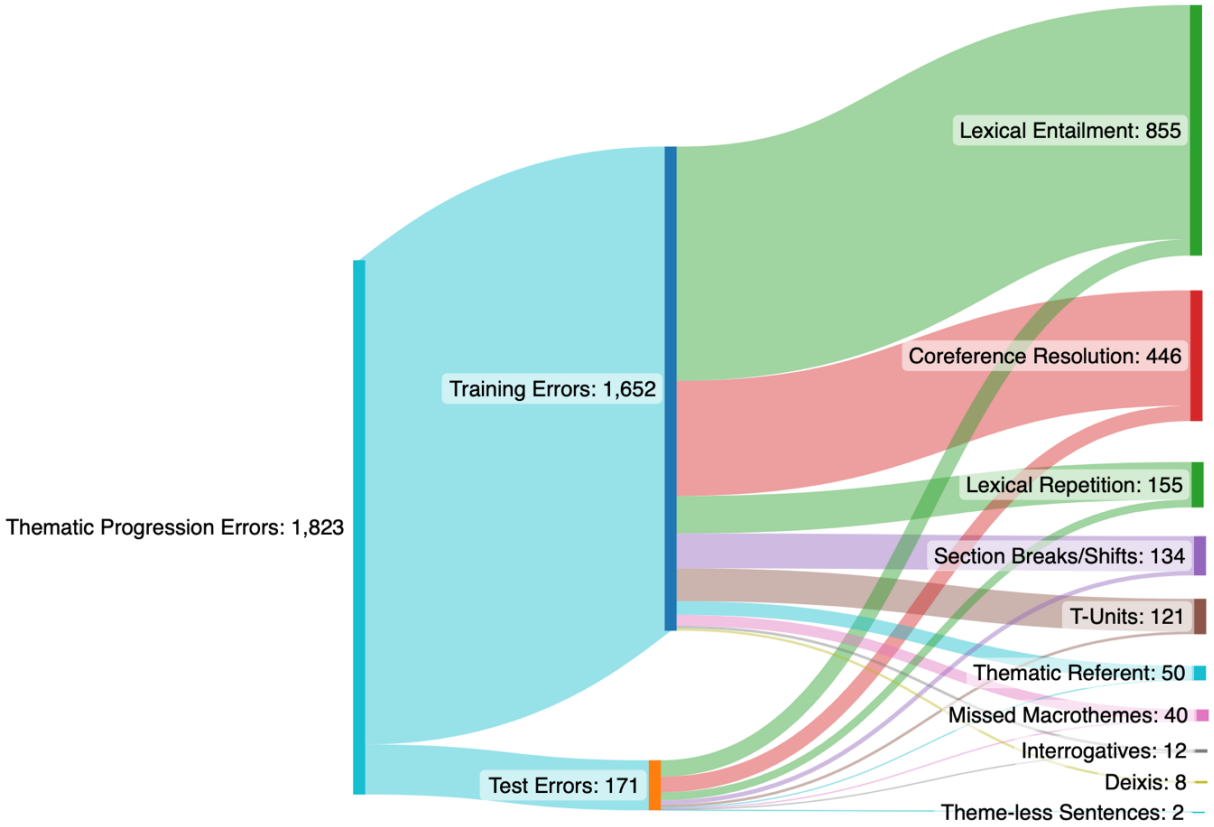


Figure 6-8: Breakdown of error cases for thematic progression classification in training and test datasets. Frequencies are provided as absolute frequencies.

As this thematic parse proved most problematic of all three through the highest frequency of errors and correspondingly lowest accuracy, each error class will be closely scrutinized in the following. A summary of the error classes to have pervaded thematic progression classification is reiterated in Figure 6-8.

Chapters 6.3.1 to 6.3.4 address the predominant error classes of lexical entailment, coreference resolution, lexical repetition and rhetorical shifts. In each section, a recapitulation of the error class definition is provided with sample misparses from Thematizer's analyses. An explanation of their emergence due to programmatic or theoretical deficiencies is provided within the context of each error class. Finally, Chapter 6.3.5 summarizes the key findings from the presented errors classes to draw final conclusions on Thematizer's ability to operationalize thematic progression.

6.3.1 Lexical Entailment Errors

Classifying thematic progression on the basis of lexical entailment caused the majority of misparses in Thematizer's third parsing task. In fact, 799 of the 1947 (41.0%) lexical entailment tests Thematizer performed were incorrect for the training datasets, 56 of the 155 (36.1%) for the test datasets. Already, this indicates Thematizer's considerable difficulty with resolving thematic progression via lexical entailment.

The main causes for Thematizer's general inability to resolve lexical entailment can be summarized into two groups: first, sub-zero similarity values as an indication of hypernymy; and second, delimiting upper and lower bounds for thematic progression pattern identification. The present section will examine these two groups of root causes to explain the effect they had on the resulting accuracy and analytical output.

First of all, as a reminder, lexical entailment covers a collection of relationships between words via hypernymy/hyponymy, meronymy, synonymy, antonymy, paraphrase and ellipsis. Hypernymy involves cases where lexical subsets of a superordinate umbrella term are employed (e.g., *a computer* → *electronics*) or where lexical items from the same semantic class are employed (e.g., *a computer* → *a router*). Meronymy indicates a part:whole relationship, whereby *foot* is meronymous to *body*. Synonymy and antonymy indicate the use of lexical items whose meaning is (generally) the same or the opposite, respectively. Paraphrase concerns the reformulation of a word, phrase or entire clause, while maintaining the same meaning as the original lexemes that are paraphrased. Finally, ellipsis as a cohesive device involves optionally omitting words or phrases or replacing these with *so* or *do* (e.g., *I like cake. So does she*). Although these classes instantiate specific lexical items, it is their semantics that determine their hierarchical and resulting syntactic relationship with one another. As such, Thematizer attempted to resolve these cases of lexical entailment through cosine similarity tests when classifying their corresponding thematic progression pattern.

Cosine similarity calculates the similarity of lexical items based on their so-called word vectors, i.e., statistical representations of a word's meaning according to a specific context. Once sentence spans are fed into the cosine similarity parser, the themes and rhemes are broken down to their content words through the removal of punctuation, grammatical terms and stop words. Afterwards, Spacy calculates the semantic similarity of the reduced themes and rhemes, which could span from -1.0 (completely unrelated) to 1.0 (identical). Thematizer finally uses this value to determine the corresponding thematic progression pattern.

Lexical Entailment Misparses due to Sub-Zero Similarity Values

Against this programmatic approach, the total 855 lexical entailment errors across both datasets (46.9% of all thematic progression errors) indicate a severe deficiency in how lexical entailment was resolved. Ultimately, the conditions imposed upon the determination of semantic values for pattern classification resulted in the high number of errors. To illuminate how Thematizer addressed lexical entailment in its parses, the sample misparses in (1) and (2) have been provided. The sentence constituents that should have instantiated the thematic progression pattern are marked in bold.

(1)	THEME	RHEME
Sentence 1	*If you commute to work, you*	*can have your car transformed into a* ***billboard*** **[R_1]** *with companies like Carvertise.*
Sentence 2	***Wraparounds*** **[T_3]**	*pay the most.*

	Thematic Pattern	Instantiation
Incorrect Thematic Progression	CONSTANT	Erroneous instantiation through thematic *you* in sentence one (S1) to thematic *wraparounds* in sentence two (S2).
Corrected Thematic Progression	LINEAR	Instantiation through hyponymy via rhematic *billboard* in S1 to thematic *wraparounds* in S2.

In (1) and in the following example (2) below, Thematizer assumed constant progression through *you* → *wraparounds* and *you* → *a hand-held spiralizer*, respectively. However, linear progression should have been identified in both cases. In (1), w*raparounds* functions as a hyponym of the hypernym (or possibly co-hyponym) *billboard*, here in the context of advertising on cars. Similarly, in (2), *a hand-held spiralizer* is an inference from *or just make your own*, such that the spiralizer can be used to *[make] veggies masquerading as pasta or rice*. Although lexical entailment can more readily account for the semantic relationship between thematic and rhematic elements in (1), resolving inferences as in (2) via cosine similarity proved particularly challenging.

(2)	THEME	RHEME
Sentence 1	*You*	*can find veggies masquerading as pasta or rice in the frozen food aisle of the grocery store, or* ***just make your own*** **[R_1]***!*
Sentence 2	***A hand-held spiralizer (for zoodles)*** **[T_3]**	*is a fun job for the kiddos.*

	Thematic Pattern	Instantiation
Incorrect Thematic Progression	CONSTANT	Erroneous instantiation through thematic *you* in S1 to thematic *hand-held spiralizer* in S2
Corrected Thematic Progression	LINEAR	Instantiation through inference via rhematic *just make your own* in S1 to thematic *a hand-held spiralizer (for zoodles)* in S2

The reason for the errors in these two examples was due to the similarity values returned upon calculation of the cosine similarity. In both cases, the constant similarity value was larger than the linear value. Specifically, the similarity value between the themes *you* and w*raparounds* in (1) yielded a value of -0.18 whereas the linear similarity value between the thematic *you* and the rhematic *can have your car transformed into a billboard with companies like Carvertise* yielded a value of -0.21. Negative similarity values between thematic and rhematic elements

across sentences in (2) were also present. Since -0.18 for constant progression was larger than -0.21 for linear progression, Thematizer parsed the pattern as constant progression.

While a sub-zero similarity value as a test condition for hypernymy resulted in correct classification of some hypernymous relationships, these examples show that it was by no means a guarantee. Proper nouns or rare words in particular, such as *wraparound* and *spiralizer*, have no pre-defined word vectors as shipped with Spacy's pipeline packages. These word vectors are required for similarity parsing as they represent the mathematical value used to express a word's contextualized meaning. The Spacy package *en_core_web_lg* that Thematizer employed for dependency and similarity parsing contains 514,000 unique vectors with 300 dimensions per vector. If no word vectors were present for the tokens' similarity parse, then an empty vector, i.e., effectively a value of zero, was used for parsing. What this means in the cases provided in (1) and (2) was that a similarity of zero for the proper nouns *wraparound* and *spiralizer* was used as a comparative value to the similarity values returned from the rhemes. This then impaired or skewed the final similarity values that the cosine similarity tests returned. Thus, similarity cosine tests that used an empty vector for similarity values contributed to the incorrect classification of thematic progression in hypernymous relationships. With cosine similarity tests specifically, the only solution would have been to provide additional word vectors through separate training for rare lexis that were not originally trained in Spacy's pipeline package.

Lexical Entailment Misparses due to Upper and Lower Similarity Bounds

A similar case whereby cosine similarity values returned the incorrect progression pattern can be seen in demonstrative pronouns within grammatical themes. These were regularly employed in the texts to paraphrase and refer back to the propositional content in the previous sentence, as shown in (3).

(3)	**THEME**	**RHEME**
Sentence 1	*In the year 2000 sex crimes*	***increased about 16.5% compared to 1980*** **[R_1]**.
Sentence 2	*The reason for* ***this development*** **[T_2]**	*could be the increasing willingness of the victims to report these crimes.*

	Thematic Pattern	**Instantiation**
Incorrect Thematic Progression	CONSTANT	Erroneous instantiation through thematic *sex crimes* in S1 to thematic *this development* in S2
Corrected Thematic Progression	LINEAR	Instantiation through paraphrase via rhematic *increased about 16.5% compared to 1980* in S1 to thematic *this development* in S2

Example (3) uses the demonstrative adjective *this* in the thematic phrase *this development* as an overt referential marker. The explicit referentiality through *this* signals to the reader that the paraphrase *this development* is a reduction of the rheme *increased about 16.5% compared to 1980*. Either the demonstrative *this* or the paraphrase *development* is optional since the presence of one element alone achieves cohesion and coherence between both sentences. The use of both, however, strengthens the syntactic and semantic connection of the sentence pair.

Rhetorically, this indicates the boiling down of previous propositional content expressed phraseologically. The NEW rhematic information from the previous sentence is reformulated as the continued foundation of the message in the subsequent theme. This approach is particularly facilitative when the rhematic information as a phrase is paraphrased in a single lexical item.

The terse demonstrative pronoun establishes immediate cohesion and coherence to the previous sentence while placing greater focus on the development of the message through the rheme.

In cases with paraphrase, demonstrative pronouns or demonstrative adjectives, Thematizer consistently classified the thematic progression pattern incorrectly. Since the paraphrase in the second sentence of (3) is derived from the rheme of the first sentence, linear progression should have been identified. In Thematizer's parse, the cosine similarity value for the concomitant themes was greater than the linear similarity value ($0.36 > 0.1$), and the constant value fell between the upper and lower bounds for constant progression. Thematizer therefore classified the progression here as constant.

The errors in Thematizer's parse indicate that it is unable to associate the demonstrative pronoun, adjective or paraphrase with its semantic partner of the previous sentence. With demonstrative pronouns specifically, the difficulty in determining the similarity between the token *this/these/that/those* and an entire phrase may appear logical from a semantic perspective. If asked about the semantic similarity between the hypothetical set *this* and *the monkey* versus the set *this* and *the apple*, it would be arbitrary to claim greater similarity of one over the other without any contextual information. Conversely, given the comparative set *this* and *the monkeys* versus the set *this* and *are playing in the trees*, incongruence between the plural *monkeys* and the singular demonstrative *this* would rule out similarity on a syntactic basis. Hence, congruence via syntactic tests could be employed across sentences to pinpoint the referent of a demonstrative pronoun.

However, when referring to entire phrases or clauses, demonstrative pronouns alone might not be a suitable candidate for cosine similarity tests in thematic progression classification. The use of demonstrative adjectives together with a paraphrase may increase the likelihood of Thematizer classifying the progression pattern correctly on account of the semantic similarity between the paraphrase and the phrase it is based on. Particularly when synonymy or antonymy were employed in the paraphrase, e.g., *changes* → *developments*, cosine similarity values yielded a more accurate representation of the sentences' propositional content. Since paraphrases can, at the surface level, appear to have little in common with previous text, contextual, cotextual and semantic contributions beneath the surface must be captured for successful coherence. In such instances, Thematizer's use of cosine similarity was able to account for this deep-level analysis to varying degrees of accuracy. A more fine-tuned and multifactorial analysis would likely be required to capture the development of propositional content in text through paraphrase or demonstrative-based thematic progression.

The same problem with demonstrative pronouns and adjective was shown to be evident in expressions that employed the elliptical *so* or as elliptical determiners, pronouns or nouns such as *some*, *someone*, *one* or *another*. In such instances, the elided word or phrase from a previous sentence was realized through an elliptical expression, whose cosine similarity was then calculated for semantic similarity tests. Just as Thematizer was consistently unable to determine similarity on the basis of previous themes and rhemes with demonstrative adjectives, the semantically void *so* and related expressions caused failure in thematic progression parses. This can be seen in (4), where the elliptical determiner *some* referred to the thematic *they* from the previous sentence.

(4)	THEME	RHEME
Sentence 1	*They* [T1]	*go by many names [...].*
Sentence 2	*Some* [T2]	*fancy themselves militias.*

	Thematic Pattern	Instantiation
Incorrect Thematic Progression	LINEAR	Erroneous instantiation through rhematic *names* in S1 to thematic *some* in S2
Corrected Thematic Progression	CONSTANT	Instantiation through ellipsis via thematic *they* in S1 to thematic *some* in S2

Since the cosine similarity value returned for the rheme *many names* and subsequent theme *some* was greater than that of *they* and *some*, Thematizer parsed this incorrectly as linear progression. Syntactically, *some* functioning as an ellipsis is logical as it represents a subset of the *many names* presented in the rheme of the first sentence. However, on the basis of the rheme that follows in the second sentence, *fancy themselves militias*, it becomes evident that *some* must refer to an animate being, i.e., *they* in the first sentence.

This example illustrates how cosine similarity can only base its output on the constituents provided as input. The answer to whether *some* was an ellipsis of *they* or *names* from the previous sentence was found in the rheme of the second sentence, which Thematizer did not account for in the cosine similarity test. Further, neither *some* nor *they* have sufficient semantic weight as they function coreferentially, i.e., their meaning is contextual and cotextual. Alone, they carry little meaning. Contrarily, *names* has greater semantic weight as it does not rely entirely on contextual or cotextual cues to manifest its meaning. While these influence the ultimate meaning *names* takes on in a text, its propositional content as opposed to grammatical or (co-)referential function affords it greater semantic weight. For that reason, the cosine similarity value between *some* and *names* outweighed that of *some* and *they*. Where tokens have greater coreferential function, therefore, such as demonstrative adjectives and pronouns or elliptical expressions, cosine similarity tests resulted in skewed similarity values.

The final example that concerns misparses of lexical entailment involves ellipsis and co-hyponyms. In (5), three sentences are presented whose thematic progression was denoted as linear and gapped linear progression. Instead, constant progression should have been identified between each of the sentences. The sentence constituents that merit constant progression have been provided in bold.

(5)	THEME	RHEME
Sentence 1	***Akodon spegazzinii*** [T1]	*is medium in size for the A. boliviensis species group.*
Sentence 2	*The coloration of* ***its upperparts*** [T2]	*varies considerably, from light to dark and from yellowish to reddish brown.*
Sentence 3	***The underparts*** [T3]	*are yellow-brown to gray.*

	Thematic Pattern	Instantiation
Incorrect Thematic Progression	LINEAR → GAPPED LINEAR	Erroneous instantiation through false thematic antecedent of *its* in S2 coreferent of rhematic *A. boliviensis species group* in S1. Further thematic progression misparse from *A. boliviensis species group* to underparts through cosine similarity.
Corrected Thematic Progression	CONSTANT → CONSTANT	Instantiation through thematic antecedent of *its* in S2 as coreferent of thematic *akodon spegazzinii* in S1. Repeated constant development from thematic *upperparts* in S2 to thematic *underparts* in S3 through co-hyponym relationship

Firstly, note the use of the singular possessive pronoun *its* in the theme of the second sentence *the coloration of its upperparts*. Here, coreference was assumed to exist between *its upperparts* and the rhematic noun phrase *A. boliviensis species group* from the first sentence. As *akodon spegazzinii* had already been identified as the theme and thereby established discourse topic, it is much more likely that the theme *akodon spegazzinii* was the coreferential antecedent instead of the rheme. This is reinforced through the specification of the coloration particular to *akodon spegazzinii* and not the generalized group. However, the erroneous coreference chain provided to Thematizer led to a linear thematic progression pattern between the first and second sentence. While this first misparse was a coreference error, it was included in the present explanation to provide further contextualization for the misparse between the second and third sentences.

There, Thematizer's failure to resolve the co-hyponymous relationship between the concomitant themes *underparts* and *upperparts* resulted in a lexical entailment misparse. Whereas constant progression should have been identified, Thematizer returned gapped linear progression. The reason for this was, yet again, the upper and lower bounds set for the respective thematic progression pattern. The similarity value between *upperparts* and *underparts* yielded 0.83, which was beyond the upper limit of 0.66 for both constant and linear progression. Upon comparing *underparts* with the rheme *varies considerably, from light to dark and from yellowish to reddish brown* from two sentences prior, the cosine similarity test yielded a value of 0.59. This fell within the pre-defined limit for gapped linear progression, which Thematizer then chose as the relevant progression pattern.

This example differs from the errors described previously in that the pre-defined upper and lower bounds for a progression pattern's similarity values precluded identification of the correct pattern entirely. The similarity value fell far beyond the similarity range for both constant and linear progression, which caused Thematizer to ignore those two patterns as possible options. While upper and lower bounds ensured that gapped patterns could be accounted for, results show that they produced false negatives too frequently. Particularly similar thematic and rhematic constituents between concomitant sentences, as was the case with co-hyponymy in (5), resulted in similarity values beyond the 0.8 mark. Such values indicate highly synonymous to nearly identical meaning, which pre-tests most commonly showed to occur in lexical repetition. As such, it was assumed that Thematizer would have caught such lexical repetitions and derivatives in the lexical repetition tests. As this was rarely the case, constant and linear patterns were consistently overlooked, and gapped patterns became overinflated.

Conclusions about Lexical Entailment

From the previous discussion, the following conclusion can be drawn concerning how Thematizer resolved lexical entailment via cosine similarity tests: The use of cosine similarity values and corresponding upper and lower bounds for thematic progression classification is an insufficient testing methodology for the determination of semantic relationships between theme and rheme constituents. Cosine similarity tests themselves were indeed able to capture more than half of the lexical entailment cases present in the texts. However, they consistently failed in cases of hypernymy, demonstratives and ellipses, whose semantic contribution could not be adequately resolved through word vector representations alone. Previous research showed similar findings, whereby cosine similarity tests alone were only able to achieve F_1 scores of 60.0% or lower (Roller et al. 2018; Agichtein et al. 2008; Saikh et al. 2015). Where research was able to achieve more reliable F_1 scores was through the use of multifactorial, distributional models that employed a pattern-based approach, i.e., "X is a (type of) Y" and embedding-based vector spaces (Vilnis & McCallum 2015; Vulić & Mrkšić 2018; Kamath et al. 2019). Instead of relying on a single deterministic value to resolve lexical entailment, contemporary models

have been trained on a multitude of lexical, syntactic and semantic factors to predict the hierarchical relationship between lexical items.

Reducing the complexity behind the semantic relationship between lexemes to a singular cosine similarity value, while convenient, shows its inadequacy when determining thematic progression. This is largely because of cosine similarity's strength in identifying how related two lexemes are in terms of antonymy or synonymy but not in terms of directionality, i.e., their hierarchical relationship. Such directionally based relationships are most prominent in hypernymy/hyponymy, meronymy and ellipsis, which is where Thematizer's parses failed most readily. Here, distributional models and pattern-based approaches as employed by the work mentioned previously would need to be incorporated for more accurate parsing. Where lexical entailment cases were successfully resolved in Thematizer's parses were then non-directional antonymy or synonymy cases, which represents the core use case for cosine similarity tests.

However, cosine similarity was not the sole reason for the high number of error cases in Thematizer's parse. The upper and lower bounds of the similarity values for differentiation between the various thematic progression patterns appeared to have hampered, rather than facilitated, classification. This became particularly evident where similarity value outliers which indicated incredibly close semantic similarity prevented Thematizer from selecting the correct progression pattern. The additional side-effect of such limits caused gapped progression patterns to become overgeneralized when similarity values far exceeded the upper bounds for constant and linear progression. In other words, higher similarity values for constant and linear progression, which should have substantiated their selection, became ignored since they were greater than their upper limit. Since Thematizer was unable to choose either of these progression patterns on the basis of exceptionally high similarity values, it defaulted to gapped progression and overinflated their presence in the texts.

While lexical entailment only represents one piece of the thematic progression puzzle, these results already reinforce the finding that Thematizer was unable to operationalize thematic progression sufficiently. The considerable number of thematic progression misparses due to unresolved lexical entailment is the first factor in Thematizer's analytical output that make the reliability of the analyses questionable. This is particularly problematic for users' interpretation and understanding of the development of their text as the misparses could invariably lead them to draw incorrect conclusions about thematic progression via lexical entailment. An expansion of Thematizer's lexical entailment parsing functionality and a re-evaluation of upper and lower bounds for cosine similarity values constitute the initial developmental direction required for greater reliability of Thematizer's output.

6.3.2 Thematic Progression Errors due to Coreference Misparses

Misclassified thematic progression patterns due to erroneous coreference resolution formed the second most common key error class in Thematizer's parses. Coreference resolution parses have been shown to impact all three thematic parsing tasks. However, thematic progression classification appears to have been affected most by this error class, constituting 446 of the 1823 (24.5%) errors. The reason for coreference misparses was either because of incorrectly identified coreference chains or because of incorrect coreference indices used when resolving coreference chains. At its core, failed coreference resolution could be traced back to incongruence between proforms and antecedents, conflation of the coreferential *it* with the dummy-*it*, and the incorrect use of antecedents from coreference chains.

Incongruence between Proforms and Antecedents

For all coreference resolution tests in the thematic classification task, Thematizer made use of the coreference indices supplied by Coreferee. Errors in Coreferee's parse invariably resulted in misparses on Thematizer's part. The first of these errors was due to the inability to correctly resolve plural proforms whose antecedent was realized in the singular.

(1)	**THEME**	**RHEME**
Sentence 1	*First, each player*	*should get a* ***player board*** **[R$_1$]**.
Sentence 2	***They*** **[T$_3$]**	*[are] double-sided.*
	Thematic Pattern	**Instantiation**
Incorrect Thematic Progression	CONSTANT	Erroneous coreference resolution through thematic proform *they* in S2 from thematic *player* in S2
Corrected Thematic Progression	LINEAR	Coreference resolution through thematic proform *they* in S2 from rhematic *player board* in S2

In (1), Thematizer should have identified linear progression due to the theme *they* in the second sentence coreferencing with the rheme *player board* from the first sentence. Instead, constant progression was returned in the parse. The rhematic *double-sided* in the second sentence substantiates the coreferential pair *they* with *player board* as, semantically, a double-sided player would be infelicitous and illogical. In such instances, Coreferee defaulted to animate antecedents if no plural inanimate antecedents could be found, which is why a coreference chain with *player* was established. Where a plural proform was used to refer to a collective noun phrase, as in (1), coreference chains were consistently false. Thematizer then used this incorrect coreference chain as the basis of thematic progression pattern classification, which explains the resulting misparse.

Coreferee's default parsing tendency toward animate antecedents can also be seen in the next example, which illustrates a greater degree of coreference complexity due to the presence of an animate, plural noun phrase in the first sentence of (2). Here, Coreferee linked the thematic possessive pronoun *their* from the second sentence with the thematic *biographers* in the first sentence. Coreferee assumed this relationship due to the congruence in number between *their* and *biographers* but also due to animacy. Since both sentence constituents were themes, Thematizer returned constant progression.

(2)	**THEME**	**RHEME**
Sentence 1	***Biographers*** **[T$_1$]**	*disagree as to the nature of the couple's relationship.*
Sentence 2	*Though their marriage was loving, some* ***biographers*** **[T$_3$]**	*suggest they viewed one another more like a brother and sister.*
	Thematic Pattern	**Instantiation**
Incorrect Instantiation	CONSTANT	Erroneous coreference resolution through thematic proform *their* in S2 from thematic *biographers* in S2.
Corrected Instantiation	CONSTANT	Repetition of *biographers* as concomitant themes should have instantiated **lexical repetition**, not coreference resolution

In fact, constant progression was the correct pattern between the two sentences; however, the coreference chain and its use as a determining factor for thematic progression in this case were

both erroneous. Firstly, the actual antecedent of *their* was *the couple*, again expressed as a collective noun phrase. While animate, *the couple* is singular, which Coreferee ignored due to the animate and plural *biographers*. Coreferee's false assumption of *their* referring to *biographers* thus resulted in the correct progression pattern but the means of progression returned was incorrect. Instead of coreference resolution, Thematizer should have returned lexical repetition due to *biographers* being realized as the theme in both sentences.

This error, while partially correct, indicates a potential flaw in the testing methodology defined for thematic progression classification. Having multiple means of thematic progression across sentences can be problematic since Thematizer performs the various thematic progression tests linearly. As a reminder, Thematizer first parses thematic progression via coreference resolution, followed by lexical repetition, macrotheme instantiation, cosine similarity and finally thematic breaks. If a test fails, then parsing moves on to the next test until thematic break is reached as a default "No thematic progression found" or "New section." This order was based on the assumption that coreference resolution would be the most frequent means of thematic progression. Results indicated, however, that lexical repetition was much more frequent than coreference resolution in both training and test datasets (49.8% and 41.9% for lexical repetition, respectively, compared to 13.6% and 12.3% for coreference resolution). Since coreference resolution was tested first in Thematizer's current version, the test for coreference resolution succeeded and the constant progression pattern was returned before moving on to lexical repetition. Therefore, the correct means of progression across both sentences could not have been identified as tests for lexical repetition were never reached in the code.

The example in (2) thereby reinforced Coreferee's tendency to default to animate, plural antecedents for the resolution of plural proforms; but it also highlighted a deficiency in the linear testing methodology in thematic progression classification. The coreference chain between *their* and *biographers* instantiated coreference resolution in Thematizer's parse, which, syntactically, was incorrect. This misparse, however, caused the correct thematic pattern to be returned, which was by no means consistently the case in coreference misparses. On account of the identified coreference chain, however, Thematizer falsely identified the means of progression as coreference resolution instead of lexical repetition, which added to the resulting misparse in the analytical output. Since lexical repetition proved considerably more frequent than coreference resolution, it should have formed the first test in thematic progression classification. This, in turn, may have prevented the increased number of coreference misparses that Thematizer yielded between both datasets.

Conflation of Coreferential *it* and Dummy-*it*

The next core reason for coreference misparses was Coreferee's inability to distinguish between the coreferential *it* and the dummy-*it* found in clefts or pleonastic structures. Similar to Coreferee's tendency towards animate antecedents in the resolution of plural proforms, coreferentiality was assumed to be resolved through the *it* and a singular antecedent from the preceding sentence, as shown in (3).

Here, Coreferee assumed that the referent of *it* was *Airbnb guest*. This was the only singular nominal phrase in the previous sentence that Coreferee could establish coreferentiality with. In actuality, the *it* functioned as a pleonastic marker akin to *It is raining*. In such instances, Thematizer should have defaulted to rhematic progression since the dummy-*it* does not offer any propositional content as the foundation of the message. Instead, the rheme is realized as GIVEN rather than the conventional NEW to form the foundation of the message. Hence, in (3), the rhematic *hits and misses* from the first sentence was developed again rhematically in the

second sentence to equate the discourse message of Airbnb visits going awry with the host's responsibility. The dummy-*it* simply functions as an emphatic rhetorical device to raise the relevance of the discourse topic introduced rhematically in the previous sentence.

(3)	**THEME**	**RHEME**
Sentence 1	*As a frequent Airbnb guest, though, there have been*	***hits and misses*** **[R$_3$]**.
Sentence 2	*And it*	***almost always comes down to the host*** **[R$_3$]**.
	Thematic Pattern	**Instantiation**
Incorrect Thematic Progression	CONSTANT	Erroneous coreference resolution through thematic proform *it* in S2 from thematic *Airbnb guest* in S2
Corrected Thematic Progression	RHEMATIC	Dummy-*it* should have instantiated default rhematic progression pattern through pleonastic structure. Further development of thematic *Airbnb guest* in S1 as rhematic *host* in S2

Since Coreferee identified coreference between *it* and *Airbnb guest*, however, Thematizer deemed the coreference test successful and returned constant progression. This then prevented Thematizer from progressing on to cosine similarity tests to identify rhematic progression. Shifting the order in which thematic classification tests are performed would not have fixed this error as would have been the case for (2). Instead, errors such as these indicate a deficiency in how Coreferee establishes coreference chains as soon as it finds proforms. This error case only applied to texts that included non-projecting clefts or pleonastic structures, which occurred much less frequently than actual instances of coreference. However, the emergence of such periphery cases highlights the continued difficulty in automatically parsing coreference while ensuring that proform approximate structures such as non-projecting clefts are ignored.

Incorrect Use of Antecedents from Coreference Chains

The final case of errors stemming from coreference misparses concerns how Thematizer made use of the antecedents from the coreference chains. The first scenario involved tracing antecedents that appeared in both the theme and the rheme of the previous sentence; the second scenario then addresses how Coreferee handled multiple coreference chains across sentence pairs. While the former revealed deficiencies in Thematizer's parsing methodology, the latter reflected historical parsing difficulties when multiple coreference chains had to be resolved through proper nouns and proforms.

Starting with repeated antecedents, such instances occurred when a proform occurred in the theme of second sentence and whose antecedents were realized in the theme and rheme of the previous sentence, as shown in (4). Here, the possessive pronoun *our* in the theme of the second sentence was resolved through the *our* in the rheme of the previous sentence. The reason for this was because Thematizer used the closest index as the antecedent. If both thematic and rhematic coreference indices were found in a chain, the closest index was invariably the rheme, which is syntactically (and indexically) closer to the theme of the second sentence.

(4)	THEME	RHEME
Sentence 1	***We*** **[T1]**	*approximated them as flat panels for our initial assessment.*
Sentence 2	*Thus,* ***our final design sketch and concept package*** **[T2]**	*seem to fulfill many of our customer requirements.*
	Thematic Pattern	**Instantiation**
Incorrect Thematic Progression	LINEAR	Erroneous coreference resolution through rhematic proform *our* in S2 from thematic proform *we* in S1
Corrected Thematic Progression	CONSTANT	Development of thematic *we* in S1 to thematic *our final design sketch and concept package* through coreference in S2

While correct, this is only part of the picture. The additional antecedent *we* was the theme of the first sentence and thereby the head of the coreference chain. This head should have been used to determine the thematic progression pattern, not the subsequent elements within the chain. If that had been done, then Thematizer would have correctly identified the pattern in (4) as constant progression instead of the incorrect linear pattern.

The reason for this misparse is a programmatic one. Once Coreferee identified all instances of coreference and the respective coreference chains, Thematizer cycled through the list of indices in accordance with the indexical span of the current sentence being analyzed. When Thematizer was fed the sentence for analysis, it compared the entire indexical span of the sentence to be processed to determine whether coreferential indices were present at all. If so, then coreference resolution took place.

S1 Indices	**0**	1	2	3	4	5	6	**7**	8	9	10
S1 Text	***We***	*approximated*	*them*	*as*	*flat*	*panels*	*for*	***our***	*initial*	*assessment*	*.*

S2 Indices	11	12	**13**	14	15	16	17	18	19	20
S2 Text	*Thus*	*,*	***our***	*final*	*design*	*sketch*	*and*	*concept*	*package*	*[...]*

Figure 6-9: Indexical breakdown of coreference chains, whose head is the thematic We *in the first sentence and subsequently realized in the possessive form as the rhematic* our *in the first sentence and thematic* our *in the second sentence. The corresponding indices in bold were used to trace the location of the coreference occurrences within the text.*

As shown in Figure 6-9, the coreference chain that Coreferee produced was the list [0, 7, 13], whereby each number represented the index of the antecedent and subsequent proforms. The textual equivalent to the indexical list was thus [*We, our, our*]; a textual and indexical representation of the coreference chain became [0: *We*; 7: *our*; 13: *our*]. During parsing, Thematizer was fed the sentence *We approximated them as flat panels for our initial assessment*. Even though the coreference index 0 for *we* and 7 for *our* fell within the sentence span [0-10], no coreference had to be resolved since *we* had no antecedent in the previous sentence. Thematizer then progressed onto sentence two, *Thus, our final design sketch and concept package […]*. Here, the proform *our* had an index of 13, which fell within the sentence span of [11-20] and it had antecedents to resolve from the previous sentence. This therefore triggered a coreference resolution test.

Once a positive hit for coreference was found, Thematizer used the previous index from the coreference chain as a search term for the previous sentence. In other words, the thematic proform *our* from sentence two formed the final element in the coreference chain; the previous

(*n-1*) index in that chain equated to *our* textually and 7 indexically in the same chain. The *our* from the previous rheme was then used as a search term to resolve the coreference from the thematic *our* in the second sentence. If the search term was found, Thematizer returned the thematic progression pattern based on the location of the antecedent in the previous sentence: if found in the theme, then constant progression; if in the rheme, then linear progression.

It is this final step that then produced the false positives in Thematizer's parse. What should have been identified is constant progression through the thematic *we* of the first sentence to the thematic *our* in the second sentence. Since *our* – and not *we* – was used as a search string, however, Thematizer only found a single instance of *our* in the previous rheme. After all, according to Thematizer, *we* did not equal *our* in terms of an identically textual, i.e., not syntactic, match between the two tokens. Therefore, relying on the textual realization of a token as a search term with the help of the *n-1* coreference index was the cause of the problem with antecedents in both the theme and rheme of the previous sentence. If the possessive proform was used rhematically, then this invariably resulted in misparses.

Where the antecedent was present in only the theme or the rheme, this problem did not occur. For example, in the sentence pair *Crawford studies linguistics. She likes books*, Coreferee identified the coreference chain indexically as [0, 4]. The pronoun *she* with index 4 had the antecedent *Crawford*, which was its *n-1* index and equated to zero in the coreference chain indices. Thus, Thematizer used the search term *Crawford*, which it found in the theme of the previous sentence. On this basis, Thematizer correctly returned constant progression.

Resolving coreference instantiated through proforms and proper nouns via the *n-1* index became additionally complicated when multiple coreference chains were identified. Such instances required Coreferee to disambiguate proforms whose antecedents were proper nouns realized within the same sentence. How coreference disambiguation was (incorrectly) resolved is shown in (5), whose coreference chains have been marked in bold. The subscripts indicate the coreference chain that Coreferee erroneously assigned the tokens to.

> (5) Rumors about amorous improprieties on **her**$_1$ husband's part affected **Virginia Poe**$_1$ so much that on **her**$_1$ deathbed **she**$_1$ claimed that **Ellet**$_2$ had murdered **her**$_2$. After **her**$_2$ death, **her**$_2$ body was eventually placed under the same memorial marker as **her**$_2$ husband's in Westminster Hall and Burying Ground in Baltimore, Maryland. Only one image of **Virginia Eliza Clemm Poe**$_1$ has been authenticated: a watercolor portrait painted several hours after **her**$_1$ death.

The text snippet revolves around Virginia Poe, who is subsequently referenced through the repeated use of *her*. However, the intermediary proper noun *Ellet* is inserted, immediately followed by another *her*. Since Coreferee often assumed the previous antecedent to resolve the anaphoric pronoun, an entirely separate coreference chain was created. Therefore, two chains emerged here: [*her, Poe, her, she, Poe, her*] for Poe and [*Ellet, her, her, her, her*] for Ellet. However, the *her* after *Ellet* actually referred to Poe and should have been included in Poe's coreference chain. Syntactically, the reflexive *herself* would have been required to substantiate a coreference chain with *Ellet*. The Ellet coreference chain was therefore superfluous and was extended until the realization of *Poe* again in the theme of the third sentence.

Following the same parsing methodology outlined in (4) and Figure 6-9, Thematizer searched for the antecedent to *her* by the time it reached the second sentence with the thematic proform *her* in *her death*. Since *her* appeared in both the theme and rheme of first sentence and was wrongly associated with *Ellet*, the thematic progression pattern was firstly misclassified as

linear progression. Instead, it should have been constant due to the thematic *her* in *her husband's part* of the first sentence.

The second error emerged between the second and third sentence. *Poe* was realized thematically in its proper noun form, which was added to the Poe coreference chain. Since the *n-1* term in the coreference chain was *she* from the first sentence, it became the search term for Thematizer to use during coreference resolution. Without *she* appearing in the second sentence, the coreference resolution test ultimately failed. If Coreferee had correctly identified the repeated *her* after *Ellet* in the first sentence as belonging to the Poe coreference chain, then Thematizer would have correctly used the search term *her* instead.

This error case thus illustrates two core deficiencies: firstly, Thematizer's misuse of the preceding index and token as a search term to resolve antecedents that appear in both the theme and rheme of the previous sentence; secondly, the difficulty behind coreference resolution when multiple coreference chains occur within and across sentence pairs. As another proper noun had been inserted between the Poe coreference chain, an entirely new yet incorrect chain was created and associated with each subsequent realization of *her*. Such cases can be particularly problematic in, but not limited to, bibliographic texts where many proper nouns may appear in succession and within a single sentence. So long as a single proper noun was realized with corresponding personal and possessive pronouns, then Coreferee largely resolved coreference correctly. This issue is not unique to Coreferee but has been a perennial issue with coreference resolution in natural language processing (cf. Stoyanov et al. 2009).

Additionally, instead of using the textual realization of the antecedent as the search term for coreference resolution, the use of the index alone would have been able to account for the various forms an antecedent may have taken on. Since the index of the antecedent already indicated its thematic or rhematic location in the previous sentence, it could have captured the thematic progression pattern regardless of textual realization. In (5), for example, switching between *Poe* and *her* should not have prevented thematic progression classification through coreference resolution on the basis of the varying coreference forms. If Thematizer had used the indices instead, then the indexical position of the antecedent could have informed classification. The use of the textual realization of the antecedent therefore indicates an overcomplication of and deficiency in Thematizer's coreference parsing methodology.

A final important note on Coreferee is its inability to account for coreference chains instantiated through the first-person personal pronoun *I* and the possessive *my*. In the sentence *My eyes filled to the brim with tears of shame. Upon being asked how I felt, I broke down crying*, Coreferee would not identify any coreference chains despite the *my* → *I* coreferentiality. While Thematizer accounted for this occasionally through lexical repetition or macrotheme instantiation, this was not always ensured and often led to misparses. Expanding Coreferee's ability to account for coreference with first-person pronouns would solve this issue. Otherwise, since this specific case was overlooked, Thematizer could have an additional syntactic parse built in that searches for such cases.

Conclusions about Coreference Resolution

Coreference resolution misparses constituting the second most frequent set of error classes in thematic progression classification indicate the severity and difficulty of tracing proforms and their antecedents as a means of thematic progression. Significant progress has been made in automated coreference resolution by computational means (see Stylianou & Vlahavas 2021, Fu et al. 2021 and Chai & Strube 2022); however, errors stemming from Coreferee parses in

assuming coreferentiality in dummy-*it* structures and inserting superfluous coreference chains through multiple entities within a single sentence highlight continued deficiencies in parsing models.
For Thematizer specifically, its parsing methodology proved deficient when leveraging the coreference chain indices for proper noun entities and their possessive pronoun form. This was shown to be particularly problematic when antecedents appeared in both the theme and the rheme of preceding sentences, whereby Thematizer defaulted to the immediately preceding antecedent and its textual realization as a search term. In doing so, Thematizer consistently assumed linear progression on account of the rhematic antecedent where constant progression should have been identified through thematic realization of the proform or proper noun.

The testing order for thematic progression was also shown to be flawed since coreference was tested before lexical repetition. As the latter instantiated thematic progression nearly three times as frequently as coreference resolution, lexical repetition should have come first in the testing order when determining thematic progression. While the coreference misparses outlined in this section would have persisted even with an altered testing order, testing for lexical repetition first could have at least reduced the number of misparses that coreference resolution caused.

On account of coreference misparses contributing to 24.5% of the thematic progression errors, further evidence for Thematizer's inability to accurately and reliably operationalize thematic progression has been found. This high frequency of errors in Thematizer's thematic progression output further complicates how users should interpret the results since falsely assumed coreference would offer a false impression of how their text was developed. Additionally, the number of false positives or false negatives in the analytical output could cause the user to draw infelicitous conclusions about how their text is thematically developed via coreference resolution. For intertextual analyses as well, erroneous thematic progression pattern frequencies on account of coreference misparses could suggest texture characteristics of certain text types that are ultimately untrue. Therefore, until changes have been made to coreference resolution so as to decrease the error frequency and increase the parsing accuracy, thematic progression output should be considered with considerable scrutiny.

6.3.3 Thematic Progression Errors due to Lexical Repetition

Classifying thematic progression by means of lexical repetition across sentence clusters was shown to be the most common means of progression in all text types, amounting to 49.8% in the training dataset and 41.9% in the test dataset. Where misparses occurred in thematic progression classification, 8.5% of them were due to Thematizer's inability to resolve lexical repetition successfully. Specifically, Thematizer failed to extract the entire noun chunk responsible for instantiating thematic progression, failed to account for different parts of speech in lexical repetition due to the unidirectional search functionality, or failed to trace acronyms realized in their complete and abbreviated form.

To recapitulate, lexical repetition was considered either the exact repetition of a singular lemma, e.g., *women*, or a complex noun phrase, e.g., *women's rights*. Tokens that belonged to the same class but were realized as a different part of speech also fell under lexical repetition, e.g., *technological* realized as the adjective of *technology*.

Failed Lexical Repetition due to Partial Noun Extraction

The first cause behind lexical repetition misparses was due to incomplete extraction of noun phrases that instantiated thematic progression. This then occasionally led Thematizer to

mismatch or overlook the entire noun phrase as the connecting element across sentence clusters. In (1), the lexical item *rural* was first instantiated in the thematic noun phrase *rural women* of the first sentence. The same lemma was then realized in the thematic noun phrase *rural areas* in the second sentence, which should have then instantiated constant progression.

(1) According to the 2011 census, the populations of **rural women** who are literate are 58.8 per cent. Still, progress in **rural areas** is delayed and often neglected.

However, as Spacy only extracted *area* instead of *rural area* in the noun phrase parse for the second sentence, the adjective was not included in the list of terms to search for repetition. The list of noun phrases extracted from the theme of the second sentence was ultimately returned as [*progress, area*]. Conversely, the list of noun phrases extracted from the theme of the first sentence was [*2011 census, population, rural women, cent*]. Thematizer then used both *progress* and *area* from the second sentence as search terms during the parse to identify any repetitions of either of these words in the first sentence. Since neither *progress* nor *area* appeared in the first sentence, Thematizer assumed no lexical repetition was present.

In the first version of Thematizer, noun phrases were foregone in favor of single lemma searches instead. This had the advantage of identifying nearly every case of lexical repetition but at the expense of overgeneralizing and missing noun phrases. Therefore, if the noun phrase *natural language processing* had been repeated across sentence clusters, then Thematizer would have only recognized *natural*, the first word of the phrase. Therefore, it would have returned with the correct thematic progression, albeit with a partial match to the connecting element. To account for noun phrases and compound nouns, then, it was decided to employ Spacy's NOUN_CHUNK iterator. This resulted in greater accuracy for determining the exact connecting elements that instantiated the thematic progression. However, partial extractions simultaneously caused false negatives in the output.

At the time of writing, it remains unclear as to why certain elements were considered part of the noun phrase and which not. As shown in (1), Spacy successfully extracted the *rural* in the compound noun *rural women* from the first sentence but not in *rural areas* from the second sentence. As such, noun phrases, including their preceding adjectives and coordination, were shown to be readily but inconsistently parsed. In Spacy's documentation, it is stated that base noun phrases alone constitute a noun chunk, which "does not permit other NPs to be nested within it" (Honnibal et al. 2020b). This would explain why the relative clause *who are literate* describing *rural women* in the first sentence was not included. The Spacy documentation does state that prepositional phrases are excluded as well so long as they are embedded at the noun phrase level. It is not certain what "embedded at the noun phrase level" means as tests with prepositional phrases as complements and adjuncts consistently returned the prepositional object as a noun phrase. Further, hyphenated prepositional phrases used as adjectives, as in *in-house patterning*, successfully returned both *house* and *patterning* as noun phrases. Finally, complex noun phrases with coordinated participles and adjectives were even returned correctly, e.g., *partially analyzed and robust parameters* correctly returned the noun phrase *partially analyzed robust parameters*. Further tests into the exact noun phrases extracted would need to be conducted to determine how Spacy specifically determines which nouns to identify as noun chunks.

Although Thematizer did check for repetition of single lexemes, the extraction of individual lemmas was based on the noun chunk parse that was originally returned. Therefore, if elements of the noun chunk were missing, such as *rural* from *rural areas*, then it was not included as a potential search term for single-lexis repetition. Extracting individual content words separate

from the noun chunk parse would have circumvented Spacy's oversight of lemmas, however. Thematizer's current inability to do so thereby indicates the first key deficiency in Thematizer's parsing methodology for lexical repetition.

Partial identification of lexical repetition was another core error when Thematizer searched for the repetition of single lexemes in preceding sentences. Such partial recognition was due to themes from the current sentence being found within previous lexical items but not functioning as a derivative form. An erroneous lexical repetition match is exemplified in (2), whereby the bold *man* was found in the token *German* two sentences prior.

(2) This section - along with section 359 - includes an exception from the double jeopardy rule, cited in article 103 paragraph 3 of the **German** Constitution. This is also known as the "ne bis in idem" principle. The rule will not allow a convict to be heard twice for the same offence. For instance, a **man** that has been arrested for theft two years ago and was acquitted, cannot be heard again for the same theft he committed two years ago.

This caused Thematizer to classify the progression pattern as gapped linear since the thematic *man* had been falsely associated with the rhematic *German* three sentences prior. Instead, it should have been linear progression via lexical entailment (paraphrase): the thematic *man* was a generalization of the more specific rheme *convict* in the preceding sentence.

This error emerged as a result of *man* partially constituting the lexeme *German* from a letter-constituent perspective. The list of nominal phrases for the theme of the final sentence was [*instance, man, theft, two years*]. None of these tokens were present when Thematizer compared them with the previous theme and rheme lists. Therefore, the next step was to investigate the presence of single-item lexical repetition at the sentence level, which took place if no repeated noun chunks were identified. Thematizer tested this condition by checking whether *man* could be found in the preceding sentences. Since the letter combination *man* was present in the letter combination *German*, the test condition was positive and Thematizer assumed a true positive match. A mismatch between *German* and *man* should have been returned on account of the two words not being identical.

In initial versions of Thematizer, Python's so-called substring search via the IN operator was not used for this exact reason. Letter combinations that appeared partially in other words invariably returned false positives. Instead, the matching expression for so-called regular expressions was used via RE.SEARCH. With this matching operation, only exact hits were returned instead of the partial matching allowance for the operator IN. Hence, with RE.SEARCH, Python would have returned a false since *man* does not equal *German*.

While regular expression afforded greater matching precision, it prevented an even greater number of single-item lexical repetitions from being identified. This was particularly the case where adjectives, adverbs and participles were employed. If the term *testing* were used in a preceding sentence and was compared with the theme *test* from the subsequent sentence, RE.SEARCH would not return a positive match. Similarly, if *temporal* were employed previously but was realized as *temporality* subsequently, only the operator IN would be able to identify the lexical repetition. Therefore, despite the invariable false positives such as those present in (2), the overall resulting accuracy for lexical repetition matching proved greater with the operator IN than RE.SEARCH. A potential improvement to the search and matching functionality would be to implement both search operations. If no noun phrases were found in the initial search phase, then single-item lexical repetition could first be queried via RE.SEARCH. Afterwards, if this second query yielded no matches, then the third, most generalized matching operation with

the operator IN could identify and extract instances of lexical repetition that are of a derived part of speech form, e.g., *temporal → temporality*. Additionally, bidirectional searching, as will be touched upon next, would also be able to account for complex → simplex patterns, i.e., *temporality → temporal*.

Failed Lexical Repetition due to Unidirectional Searching Methodology

Thematizer's occasional failure to recognize lexical repetition through a different part of speech or simply different form of the word impacted tracing single-word repetition specifically. This was largely due to the unidirectionality of the search parse that Thematizer performs, as illustrated in (3). Here, a compound misparse occurred, whereby coreference through the coreference chain *Vegan Lilac Lemon Cake* and *it* remained unresolved due the contracted *it's*. As both *Lemon Cake* and *it* form the themes of both sentences, coreference resolution would have been the correct means of progression to substantiate the constant progression. Otherwise, the realization of the thematic *lemon* in the first sentence as thematic *lemony* in the second failed to return a positive hit.

(3) This Vegan Lilac **Lemon** Cake is the most flavorful and delightful cake for spring! **Lemony** and bright, light and tender yet buttery, it's the best of both worlds.

The issue Thematizer had here was the order in which the elements are searched for. Upon extracting noun phrases, the list of search terms from the second sentence was returned as [*lemony, bright, light, tender, buttery*]. Thematizer then used each token individually as a search term to check for its realization in the preceding sentence. Hence, *lemony* was compared with the tokens in *Vegan Lilac Lemon Cake*, which returned a false, since *lemony* was not present in the theme. Then, *lemony* was compared to the rheme *is the most flavorful and delightful cake for spring*, which again returned a false. Since *lemony* was not found in the preceding theme or rheme, Thematizer then continued with the next item in the list *bright* and repeated the process until all search terms were used.

When comparing *lemony* with the previous theme containing *Lemon*, a partial match seems evident: the letter *y* has simply been added to *lemon* from the theme in the first sentence. However, in plain terms, the matching test that Thematizer performs is whether the exact letter combination *lemony* is present within the phrase *Vegan Lilac Lemon Cake*. Since there was a *y* at the end of *lemony*, Thematizer concluded that *lemony* does not equal *lemon* and no lexical repetition is present.

The reason for this is how Thematizer ascertains whether a token is present in another phrase. The individual letters that comprise *lemon – l, e, m, o* and *n* – are, in that order and as a single constituent, present in the token *lemony*. For Thematizer, it is irrelevant that a *y* is at the end of the phrase because the core condition of the five letters in *lemon* has already been identified. Conversely, checking whether each individual letter in *lemony* are present in *lemon* is false since *lemon* lacks a *y*. Since Thematizer only uses the tokens from the current theme for comparison against the tokens of the themes and rhemes from preceding sentences, it only searches in one direction: backwards. In other words, Thematizer only used *lemony* and searched backwards to check for lexical repetition. Had it also searched for lexical repetition using tokens from the preceding sentence to the current theme, it would have found *lemon* subsequently realized as *lemony*.

All search terms were lemmatized in order to ensure greater uniformity in the form of the tokens to be searched. With this pre-processing step, it was assumed that unidirectional searching

would suffice as the base form of tokens would be compared against one another. However, as the lemmatized form of *lemony* was itself *lemony*, this shows that some cases were able to circumvent a uniform structuring of search tokens through lemmatization. A lack of bidirectional search parsing thereby indicates the second key parsing deficiency in Thematizer's methodology for lexical repetition.

Failed Lexical Repetition due to Acronyms

The last issue to have contributed most frequently to errors in the lexical repetition tests was Thematizer's failure to trace instantiations of acronyms in their abbreviated and full form. Once proper nouns were first introduced with their acronym in parentheses, the acronym alone was used in subsequent sentences to refer to the entity. In (4), the two entities *General Electric Company* and *Marconi Electronic Systems* are realized in their full form initially and subsequently referenced as their respective acronyms *GEC* and *MES*.

(4) Between 1945 and 1999, **GEC-Marconi/Marconi Electronic Systems** became one of the world's most important defence contractors. **GEC's** major defence related acquisitions included Associated Electrical Industries in 1967, Yarrow Shipbuilders in 1985, Plessey companies in 1989, parts of Ferranti's defence business in 1990, the rump of Ferranti when it went into receivership in 1993/1994, Vickers Shipbuilding and Engineering in 1995 and Kværner Govan in 1999. In June 1998, **MES** acquired Tracor, a major American defence contractor, for £830 million.

Since *GEC* was used as an acronym in both the first and second sentence, Thematizer was able to successfully identify it as the repeated lexical item that instantiated constant progression. However, between the second and third sentence, Thematizer failed to associate *MES* with *Marconi Electronic Systems*, which should have been identified as gapped constant progression. Since no lexical repetition was identified, Thematizer progressed on to the next thematic progression test, macrothemes. If acronyms were used frequently enough in the text to achieve discourse relevance through Latent Dirichlet Allocation, it then became a macrotheme. In that case, Thematizer then would have identified *MES* as a macrotheme. In (4), this did not happen, however, so Thematizer had to resolve thematic progression via cosine similarity, which ultimately returned gapped constant progression. Therefore, while the progression pattern returned was correct, the means of progression should have been lexical repetition, not lexical entailment.

This breakdown indicates how Thematizer was able to account for acronyms' contribution to thematic progression while simultaneously being unable to identify the full form of an acronym in the text. As long as the abbreviated form of a compound proper noun was absent from a text, Thematizer failed to identify the lexical repetition that the acronym instantiated.

Even if an acronym was provided, however, this information was lost when cleaning the text during pre-processing. Before thematic analysis, the text was cleaned of all parentheses and brackets, including any information found within these punctuation marks, as the information therein was typically noise in the form of citations, years or anecdotes. Information that was not noise but removed nonetheless was acronyms in parentheses after they had been initially mentioned in their full form. As such, the first instance of the required acronym that Thematizer could have used for lexical repetition was deleted, as was the case in (4).

While readers can associate acronyms with their referents even without subsequent repetition, doing so computationally would have required making acronyms from all proper or compound

nouns from a text for potential realizations as lexical repetition. This would have not only added considerable computational overhead during pre-parsing but also increased the likelihood for false positives given the number of permutations that proper and compound nouns in a text could provide. Since acronyms were only deleted when found between parentheses but maintained otherwise, excluding their removal during text cleaning would ensure Thematizer's ability to identify lexical repetition via acronyms more readily.

Conclusions about Lexical Repetition

The key error class of lexical repetition misparses causing misclassification of thematic progression patterns was shown to be due to partial noun phrase extraction, failure to account for a lexeme's varying realizational form through unidirectional search parsing, and finally, acronyms. This group of errors is among the first of the key error classes to constitute less than 10.0% of all the thematic progression classification misparses. Of the 4396 cases of lexical repetition found in both datasets, only 155 or 3.5% were actually incorrect. This indicates that Thematizer was largely able to account for thematic progression via lexical repetition despite the total 155 error cases. While deficiencies were identified in Thematizer's parsing methodology, particularly in identifying certain cases of single-lexeme repetition and in the lack of bidirectional searching, the programmatic approach to lexical repetition ultimately proved robust. In particular consideration of the high frequency of lexical repetition, a greater number of errors would have been made manifest if the underlying parsing methodology had had fundamental flaws.

It can therefore be concluded that Thematizer was able to accurately and reliably trace thematic progression when instantiated through lexical repetition. Aside from macrotheme instantiation, lexical repetition proved to be most successful in terms of accurate parses when classifying thematic progression. The errors outlined in the present section have revealed residual deficiencies in specific thematic parses, but their relative infrequency suggest their secondary or peripheral emergence.

The high accuracy of thematic classification via lexical repetition was likely on account of the comparative simplicity in its parse: Whereas lexical entailment and coreference resolution relied on semantic and syntactic tests that were complicated through numerous contextual, cotextual and dependency factors, lexical repetition revolved solely around the identification of lexemes in text. So long as identical lemmas or their related forms were successfully identified through repetition across sentence clusters, then the corresponding thematic progression classification could readily be ascertained. By checking lexical repetition across concomitant themes first as well, Thematizer was able to maintain the development of previously established discourse topics over those that had been introduced in the preceding rheme for the first time. In doing so, appropriate precedence to the development of the foundation of the message was accurately given while accounting for the myriad thematic progression patterns that were instantiated through lexical repetition.

6.3.4 Thematic Progression Errors due to New Sections and Rhetorical Shifts

Both rhematic progression and thematic breaks were two patterns that can indicate a shift to new topics or rhetorical sections. Otherwise, they may reflect a marked break from the typical GIVEN-NEW informational structure to emphasize the propositional content of the sentence in question. Both thereby served as marked rhetorical devices in the structural and rhetorical development of a text.

As presented in Chapter 5.4, thematic breaks were shown to either be frequently overlooked or overgeneralized as a default structure once all other thematic progression tests failed. While the latter was an intended functionality to account for potential thematic breaks, parsing errors that emerged during thematic classification often resulted in thematic breaks being superfluously identified. Rhematic progression was also shown to be frequently overlooked, which may have been due to its infrequent occurrence. Ultimately, the reasons for misparses with thematic breaks and rhematic progression as indicators of section breaks or shifts were: the linearity in testing methodology for these two patterns; discourse markers introducing new rhetorical sections; and Thematizer defaulting to thematic breaks due to cascading errors from pre-processing and the index identification task.

Linear Testing Methodology as Cause behind Erroneous Identification of Section Shifts and Breaks

In the identification of thematic breaks and new rhetorical sections, the role that NEW topics play is pivotal. With thematic breaks specifically, NEW topics are defined as ones that have not been previously employed in the text but are nonetheless realized thematically. Themes introducing NEW discourse topics transfer the GIVEN informational status to the rheme, where previously established discourse topics are then realized. Hence, information status is deliberately switched for emphatic, stylistic or rhetorical purposes. When GIVEN discourse topics are developed across two concomitant rhemes, rhematic progression is at hand. When progressing onto a new rhetorical section of a text, there may not necessarily be a thematic connection to the previous sentence. In such instances, a NEW topic derived from overarching discourse topics may be realized thematically to introduce a conscious thematic break from the previous sentence and section. An overt example of this could be the progression from one chapter to the next in a book. Less overt examples are present between sections of a report or even between paragraphs.

Whereas Thematizer was more readily able to identify overt thematic breaks and rhematic progression due to small cosine similarity values, less overt instances proved problematic. The examples in (1) and (2) both illustrate rhematic progression through the insertion of a new topic as the theme (in bold) in the second sentence. There, the rheme exceptionally takes on GIVEN information status as a development of a previously established discourse topic. The <§> mark indicates the shift to rhematic progression through the introduction of a NEW discourse topic.

(1) **Ultimately, Stevenson is arguing that the education of moral attitudes** offers no other reason past the statement, “Because I said so.” <§> **Few children** have been satisfied with such a justification.

(2) **Because the United States government is so busy promoting and spreading democracy, it** leaves little room for criticism. <§> **I** am interested in a critical look at specific cases of transition to democracy, concentrating on the problems associated with these transitions, with the intention of improving the contemporary practice of democracy as it is applied in different situations.

In both examples, Thematizer assumed constant progression on account of the semantic similarity values between the concomitant themes being greater than linear progression values. Considering the individual themes in bold, their realization cannot be derived from either the theme or the rheme from the previous sentences. For example, *few children* cannot be seen as a development of *Ultimately, Steven is arguing that the education of moral attitudes* or its rheme through coreference resolution, lexical repetition or macrotheme instantiation. Further, there is

no hypernymy, antonymy, paraphrase or ellipsis between the thematic and rhematic constituents compared to the second theme's *few children*. Similarly, the thematic *I* in (2) cannot be derived from any constituents of the previous sentence.

Instead, the connecting elements are found across the rhemes of both examples: *Because I said so* in (1) is developed through the paraphrase *such a justification* whereas *democracy* is developed through lexical repetition as *democracy* again in (2). Thematizer's failure to recognize rhematic progression here was due to where it was tested in the parse, namely after constant and linear progression. If cosine similarity tests yielded a hit for constant or linear progression, then rhematic progression was not tested at all. Since Thematizer indeed returned constant progression for both of these examples, its identification prevented rhematic progression from being tested entirely.

The decision to test rhematic progression last was because of its relative infrequency in text. Rhematic progression only occurred collectively in 5.6% of both datasets' thematic progression patterns. Further, rhematic similarity values were larger than both constant and linear similarity values in more than 97.0% of the cases. If rhematic had been tested first, then constant and linear would have, in turn, been the overlooked pattern. Thus, linearity in testing again posed a problem when similarity values for the three patterns – constant, linear and rhematic – fell within the pre-defined bounds. Since constant and linear progression were the most common patterns in the datasets (34.2% and 26.0%, respectively), they were tested for first to appropriately account for their prevalence and contribution to text development. The only exception to this was clefts and existentials, which were unique cases that Thematizer searched for initially. If found, they circumvented constant and linear progression tests and automatically defaulted to rhematic progression.

A similar case, albeit with a thematic break, can be seen in (3), where neither thematic nor rhematic constituents of either sentence develop the discourse message across the sentence pair. The break here is indicated by the <||> symbol, and themes are given in bold.

(3) **It** functions really well as a big house with lots of people rattling around in it. <||> **On a beloved hobby: Milly** embroiders pillows all the time.

Here, the circumstantial theme *On a beloved hobby* acts as an overt section header and indicator of a new rhetorical section. The cosine similarity test between the rheme of the first sentence and the theme of the second yielded a value of 0.36. This was greater than the constant progression similarity value and thereby prompted Thematizer to identify the progression pattern as linear progression. However, due to lack of implicit and explicit connecting elements – semantically, lexically or cohesively – linear progression cannot be justified. While one could claim distant lexical collocation between *house* and *pillows*, it would likely not be strong enough to substantiate rhematic progression across the two sentences. Therefore, since there were no connecting elements present, Thematizer should have identified a thematic break instantiated through a new rhetorical section upon failing the previous thematic progression tests.

In practice, however, there are two arguments against the approach to defaulting to a thematic break when previous thematic progression tests fail. First of all, this approach assumes that the syntactic and semantic parses done beforehand were perfect. However, false negatives in the previous parses resulted in false positives for thematic breaks. In other words, if coreference resolution, lexical repetition, macrotheme instantiation and (gapped) constant or linear progression were not identified but should have been, Thematizer would return a thematic break.

Previous discussions have outlined deficiencies in coreference resolution and the unreliability of cosine similarity for lexical entailment. Errors also emerged during pre-processing, such as incorrect t-units, which complicated subsequent thematic progression parses even more. As such, a defaulted thematic break may have been falsely identified on account of misparses from tests occurring before thematic break identification.

The second argument against defaulting to thematic breaks when all other tests fail is that new rhetorical sections may be present despite the success of other thematic progression tests. Consider the following fictitious example whose themes are in bold and whose break is indicated by <||>:

(4) **Chowder** is a big hit in my family, and **I** got the recipe for it from my great-grandmother. **It** uses fresh vegetables and nice autumnal spices, perfect for a cold and rainy October day. **This chowder** goes well with green salads, sandwiches and quesadillas, and simple vegetable sides. <||> **A few nice (salad) options** are as follows: Firstly, […]

Here, the final sentence was written such that *a few nice options* in the final sentence could be understood as a paraphrase of the preceding rheme. The term *salad* in parentheses could be added as well, which would have caused Thematizer to recognize the development as linear progression through lexical repetition as well. A new rhetorical section is overtly introduced through the shift to the new discourse topic *salads*, the phrase *are as follows* and the subsequent listing of the options. In doing so, the predominant discourse topic of *chowder*, evident in every sentence but the last, shifts to *salad* in the final sentence. Hence, within this paragraph, a microstructure appears whose rhetorical sectioning is instantiated through the overt lexical repetition of *salad* and *are as follows* phrase as a distinct divergence from the previously established topic *chowder*.

With how Thematizer parses thematic progression, however, a shift in the rhetorical section would have been overlooked since lexical entailment or lexical repetition would have accounted for the means of progress across each sentence. Thematizer's general failure to identify thematic breaks due to the success of previous thematic progression tests may impair the impression of how a text is structured around the treatment of primary discourse topics. While its attempt to find progression between each sentence can trace discourse development at the micro (sentence) level, it was at the macro (paragraph) level where Thematizer missed thematic breaks most. Thus, more subtle shifts in rhetorical sections were insufficiently accounted for where a change in discourse topic could have substantiated a thematic break. It should be noted that macrothemes were tracked for this specific reason as well: where thematic breaks remained unaccounted for, macrothemes often accounted for shifts to new discourse topics that were relevant at the document or paragraph level.

One would otherwise arguably assume that thematic breaks could have been substantiated through incredibly low, negative similarity values. After all, the smaller the similarity value, the less related the compared elements are. However, as Thematizer was programmed to reserve negative similarity values for cases of hypernymy and antonymy, thematic breaks could not be resolved via such values. Additionally, instances of thematic breaks often returned positive similarity values that fell within the upper and lower bounds for (gapped) constant and linear values. Therefore, adding upper and lower bounds to thematic breaks would have had no effect. In fact, as discussed in Chapter 6.1, imposing limits to the similarity values for thematic breaks would have likely been of greater detriment than of benefit.

That being said, just as the linear testing order in cosine similarity tests prevented rhematic progression from being identified, defaulting to thematic break after all other tests failed proved to be a deficient approach. In the first versions of Thematizer, cosine similarity tests were performed for every thematic progression pattern across sentence clusters. Whichever pattern achieved the highest cosine similarity value was then selected as most appropriate. Not only did this overinflate rhematic and gapped progression patterns, it exponentially increased the processing time required for thematic progression classification. For that reason, a linear approach was decided on, but with the unintended side effect of rhematic progression and thematic breaks being unduly accounted for.

A final minor case to have emerged is how discourse deixis were parsed with cosine similarity tests but should have been automatically marked as a thematic break. While the use of discourse deixis through *here*, *there* or the extratextual *it* and *that* was incredibly rare (0.5%), it was consistently unaccounted for.

(5) **This chowder** goes well with green salads, sandwiches and quesadillas, and simple vegetable sides. <||> **Here** are a few nice options.

In (5), there is no grammatical theme since the grammatical subject *options* follows the verbal root *are*. Instead, only the circumstantial theme *here* is present. The lack of a grammatical theme should have forced Thematizer to default to a thematic break or rhematic progression. However, since the circumstantial theme was extracted during cosine similarity parsing, it was used for comparing word vectors to return linear progression.

The use of the cataphoric *here* signals the introduction of a new rhetorical section that Thematizer overlooked. One could argue for rhematic progression between both sentences as the rheme from the first sentence *goes well with green salads, sandwiches and quesadillas, and simple vegetable sides* is elided in the rheme of the second sentence, i.e., *a few nice options [for green salads, sandwiches and quesadillas, and simple vegetable sides]*. However, as the text continues with an elaboration of said *few nice options*, beyond *green salads, sandwiches*, etc., the entirety of the second sentence becomes NEW through the fronted *here* and lack of connecting elements to preceding themes or rhemes. In fact, this is similar to the use of existentials, whereby new discourse topics are overtly introduced. However, while existentials do not necessarily introduce a new rhetorical section, the use of *here* can. As the use of such discourse deixis was both rare and exceptional in its function, Thematizer could be programmatically equipped with an exception case to address such overt instantiations of thematic breaks.

Thematic Breaks as a Result of Cascading Misparses

Aside from thematic break and rhematic progression being overlooked due to the linear testing methodology to thematic progression, Thematizer's failure to identify a grammatical subject and/or root resulted in default thematic breaks. This was either the result of erroneous dependency parses or pre-processing errors, such as t-unit splitting or direct quotation parsing. Just as cascading errors affected marked theme classification, they resulted in the frequent misparse of thematic breaks during thematic progression classification. Example (6) illustrates how Thematizer did not split the direct quotation from the projecting matrix clause *she says*, which thereby prevented identification of the grammatical themes in both clauses. Themes that should have been identified are in bold.

(6) Choose one day per week that you will do your grocery shopping, and any other shopping, Doe suggests. "**This** will help you practice money mindfulness and eliminate impulse spending," **she** says.

In the first sentence, the grammatical subject, *Doe*, appeared in the projecting matrix clause after the projected imperative. Since it appeared after the imperative, the entirety of the sentence became a thematic break as no lexical repetition was present in the preceding rheme to instantiate rhematic progression. The same structure was found in the second sentence, which Thematizer should have split into two t-units: one t-unit for the entire direct quotation and one for the projecting matrix clause *she says*. As Thematizer was unable to identify either *this* or *she* as the grammatical subjects of their respective sentences, it did not split the direct quotation from *she says*. While the cause of the misparse was Thematizer failing to split the direct quote from the projecting clause, the two-fold effect it had on the resulting parse was its inability to identify the grammatical theme and the misclassification of the thematic progression pattern.

As has already been shown with a number of t-unit misparses in the previous thematic parsing tasks, failure to split t-units and other pre-processing errors had cascading effects and manifested as thematic break errors in thematic progression classification. Therefore, the key error class of thematic breaks emerging due to dependency misparses was a direct result of t-unit misparses from the beginning of the thematic analyses. Further, incorrectly splitting a single independent sentence into multiple t-units caused Thematizer to introduce thematic progression patterns that were neither present in the original nor did they need to be accounted for.

The cascading errors due to dependency and pre-processing errors totaled 121 cases for t-unit misparses and secondarily caused nearly half of all thematic break misparses. This reinforces the far-reaching consequences that misparses early on in the thematic analysis can have on residual parses. In light of the 1198 cases of thematic breaks found in the training and test datasets, an error frequency of 121 or 10.1% indicates how pervasive cascading errors were in thematic progression classification. As touched upon in the discussion of the previous thematic parsing tasks, addressing the root of these errors would invariably have a positive cascading effect upon their resolution in pre-processing and the index identification task.

Conclusions about New Sections and Rhetorical Shifts

The findings on how Thematizer identified or overlooked shifts in rhetoric and structural sections via rhematic progression and thematic breaks have shown to indicate an additional stumbling block in the parsing methodology. An error rate of 14.5% (134 total errors stemming from new sections/rhetorical shifts, 121 errors from t-units, 8 errors from deixis and 2 from sentences without a theme, given the total 1823 thematic progression errors) indicates that Thematizer was able to account for such shifts in the majority of the cases. However, it leaves a large portion of implicit and explicit shifts in discourse unaccounted for. Testing for rhematic progression after constant and linear cosine similarity tests led to Thematizer most frequently missing the marked rhetorical device of shifting GIVEN and NEW information status for emphatic purposes. Further, relegating thematic breaks to a default catch-all test case upon failure of all other thematic progression tests caused Thematizer to assume thematic breaks superfluously. Errors in coreference resolution, t-unit parsing, deixis and sentences without a theme further compounded the issue by shifting unsuccessful resolution of these test cases to a presumed thematic break.

As touched upon in Chapter 6.3.1, the removal – or at least further scrutiny – of the upper and lower bounds for thematic progression identification merits consideration with particular respect to rhematic progression. Since cosine similarity values between concomitant rhemes were consistently and significantly greater than those of constant and linear patterns, this puts into question whether similarity values are an appropriate measure for rhematic progression identification. The development of the discourse message across themes is exceptional in and of itself as it contradicts the core tenet of thematic progression, whereby established discourse topics are generally realized thematically. If this rule is deliberately broken for rhetorical or stylistic reasons and shows statistically significant frequency in its use, it is worth examining the factors behind its exceptional instantiation.

While previous work has examined the presence of rhematic progression in various text types (cf. Leong 2005; Hawes 2001; Li 2009), little discussion has been dedicated to the emergent factors behind the pattern. This may have been due to its relative infrequency across text types but also due to thematic theory itself: The thematic paradigm places emphasis on the development of previously established discourse topics realized thematically in text. While elements derived from rhemes are core constituents of the thematic model, they are always contextualized as subsequently realized themes, i.e., newly established discourse topics. After all, as Fries stated, it is the experiential and thematic elements that are responsible for the foundational discourse structure of a text (1995: 319). Rhemes, on the other hand, are then subsumed under the generalized category of constituents that develop the theme and thereby the discourse message. While their importance is not neglected in a thematic analysis of text, it has historically been given little attention.

For Thematizer and the present work, cosine similarity tests with upper and lower bounds have shown to be an insufficient method for identifying rhematic progression. As it is the shift in GIVEN-NEW information structure that makes rhematic progression exceptional, this should have formed the basis of its automated identification. Tracing which discourse topics have already been established within a text and where they are subsequently realized in a sentence – either thematically or rhematically – could serve as an initial stepping stone for the targeted identification of rhematic progression. Research focusing explicitly on the propositional content, syntactic realization and contextual factors of rhemes could further delineate the means by which rhematic progression could be operationalized and computationally analyzed.

As for thematic breaks, fine-tuning the parameters required for its successful identification may ameliorate its overgeneralization in Thematizer's current version. To date, most research on the investigation of thematic breaks has focused on cohesion measures to determine whether a break in the propositional and structural development of text has occurred (Shum et al. 2017; Crossley et al. 2013; Ferret & Grau 1998). Instead of or in addition to relying on the fail cases of previous thematic progression tests, breaks could be quantifiably justified by determining the degree of cohesion between sentence clusters. This would offer an additional parsing parameter for thematic break tests that would facilitate an isolated requirement for their identification.

Otherwise, rhetorical structure models have also been employed to trace implicit and explicit breaks as a means to indicate where structural breaks in discourse occur (Simsek et al. 2013; Joty et al. 2015). With the use of probabilistic models, sentences could be reduced to discourse units and parsed in terms of their entropy as a function of their rhetorical function at the sentence and document level. The added degree of parsing complexity would enable a more fine-grained analysis of the rhetorical structure a text has, albeit with considerable computational overhead added to Thematizer's parsing functionality.

Both approaches provide quantifiable and theoretical frameworks for a comprehensive analysis of discourse scaffolding, which could inform future versions of Thematizer's parsing methodology. Compared to the deployment of rhetorical structure models, cohesion analyses would be a programmatically and conceptually simpler approach. In this simplicity, the resulting analyses could lack the necessary details for readily capturing thematic breaks and rhetorical shifts. This also runs the risk of belying the complexity inherent to the coherent and cohesive structuring of texts that rhetorical structure models could more accurately capture. Future research and changes to Thematizer could investigate the benefits each approach offers to determine which methodology would remedy the residual 14.5% errors from Thematizer's current programmatic approach.

On the basis of these errors and the findings outlined in the present section, Thematizer's operationalization of thematic structure in terms of shifts in rhetorical sections and thematic breaks proved insufficient. The average F_1 score for thematic break parses amounted to 71.2%, which is 8.0% less than the gold standard for thematic progression analyses. While this indicates that most thematic breaks were successfully accounted for, the large margin of error makes Thematizer's analytical output questionable. Therefore, similar to lexical entailment and coreference resolution, it can be concluded that rhematic progression and thematic breaks also fall under the cases of inconsistent and deficient operationalization with Thematizer's current parsing methodology.

6.3.5 Key Takeaways from Thematic Progression Classification

In summary of the key error classes and their root causes for thematic progression, results have confirmed that Thematizer's operationalization of thematic progression was insufficient. Where Thematizer struggled most was with resolving lexical entailment, coreference, capturing rhetorical shifts via thematic breaks and, partially, lexical repetition. Altogether, these accounted for 1566 of the 1823 (85.9%) errors in thematic progression classification and severely impacted the resulting accuracy in both training and test datasets. Considerable adjustments to Thematizer's parsing methodology, particularly to that of lexical entailment tests, would be required to reduce the occurrence of misparses. That being said, Thematic progression classification via lexical repetition (including macrothemes) was shown to be more reliable despite the number of errors that emerged.

Thematizer's parsing accuracy was thus severely impacted through the pervasiveness of the key error classes outlined throughout Chapter 6.3. Lexical entailment remained largely misparsed due to pre-defined upper and lower bounds for each progression pattern. As a result, where Thematizer could have accounted for semantic relationships across sentence clusters, these bounds prevented correct classification. This issue was compounded with demonstrative pronouns and adjectives, which served a paraphrasing, coreferential function to propositional content in previous sentences.

Next, coreference resolution misparses partially resulted from Coreferee identifying the incorrect antecedent. Thematizer used the indices that Coreferee marked as coreferential, which proved problematic when proforms were mismatched or entirely ignored with first-person personal pronouns. Pleonastic expressions realized through the dummy-*it* in clefts were frequently misparsed despite Coreferee's embedded ability to account for these. Finally, using the textual realization of proform-antecedent pairs instead of their indices caused Thematizer to overlook or misidentify how text was developed thematically via coreference resolution.

Partial extraction of noun phrases, particularly due to the use of acronyms, led to Thematizer misclassifying the progression pattern in lexical repetition tests. Spacy's inconsistent identification of noun chunks and its dependent constituents failed to return the entirety of lexical phrases. This, in turn, prevented Thematizer from employing the relevant tokens as search terms during the matching parses. The repetition of lexical items using a different part of speech also led to failure cases despite lemmatization of search terms. This was further complicated by matching search terms unidirectionally, i.e., from the current theme to preceding themes and rhemes alone, instead of bidirectionally. The use of the Python operator IN instead of the more limiting RE.SEARCH returned overgeneralized cases of lexical repetition. Similarly, misidentification of the connecting elements responsible for instantiating thematic progression via lexical repetition yielded mixed results: The corresponding thematic progression pattern was successfully returned in most cases, but either part of the repeated phrase was overlooked, or entirely different tokens were assumed to instantiate progression.

Thematizer's tendency to overlook thematic breaks and new discourse topics through rhematic progression proved to be an additional parsing hurdle. New discourse topics were overlooked due to cosine similarity values between thematic and rhematic elements, which invariably prevented Thematizer from recognizing rhematic progression. Here, the order in which progression patterns were tested contributed to the oversight in thematic breaks' presence. Failure to correctly identify the grammatical subject on account of compound independent clauses or direct quotes caused Thematizer to assume a thematic break or new topic where none was present. Finally, the use of particular deictic phrases as a signal for a new rhetorical section remained overlooked.

Cascading errors stemming from the index identification task affected thematic progression classification through the introduction of new sentences that were dependent in nature and therefore had no thematic progression. How Thematizer tested for independence as grounds for splitting compound sentences into multiple t-units was the cause of the misparses. Most commonly, independent clauses that functioned within an embedded, dependent structure, e.g., direct interrogatives as object of the preposition, and direct quotations caused Thematizer to assume independence and substantiate a split.

While a complete rework of thematic progression classification may not be necessary, closer examination of how similarity tests are conducted merit reconsideration. Fine-tuning coreference resolution with the results Coreferee provides and how Thematizer handles the coreferencing indices would further aid in overall accuracy improvement. Residual issues with lexical repetition through implementing bidirectional search patterning, as outlined in Chapter 6.3.3, could then boost Thematizer's performance even further. Additionally, the linear texting order for thematic progression requires further examination in order to account for instances where multiple progression initiators exist across sentences. The exact changes that could be made to these parsing tests thereby forms the foundation of future work and version of the program.

The ultimate conclusion to be drawn from the findings and key errors cases outlined in Chapter 6.3 is that Thematizer was unable to operationalize thematic progression. The high occurrence of errors in Thematizer's parses impaired its parsing accuracy to such a degree that the reliability of analytical output must be questioned. Therefore, writers' accessibility to thematic structure via thematic progression specifically remains unattained.

6.4 Final Conclusions about Thematizer as an Automated Feedback Tool for Writing

This final section seeks to recapitulate the key findings presented throughout Chapter 6 as contextualized around Thematizer's functionality, output and purpose. The development of Thematizer was driven by improvements upon previous computational and theoretical approaches to thematic analysis. In doing so, the development of an automated text analysis tool was targeted so as to make thematic theory accessible to writers. How well the present research, and thereby Thematizer, was able to achieve these goals will be outlined in the following in terms of its parsing functionalities, analytical output and resulting deficiencies.

The greatest caveat to Thematizer's automated output was the errors that were inevitably part of the results. Marked theme classification proved to be the sole parsing task to meet or exceed the gold standard from previous research. While Thematizer was able to achieve accuracy rates near or beyond the gold standard for the index identification and thematic progression tasks, neither was perfect; low accuracy rates for the thematic progression classification task were particularly problematic due to its 80.2% and 75.9% F_1 scores for training and test datasets, respectively. Addressing the underlying causes for misparses in index identification and thematic progression classification in particular would be required for the output to be considered reliable. Until Thematizer is able to consistently exceed gold standard F_1 scores in its first and third parsing task, manual correction of the analytical results still proves necessary. For that reason, while text analysis with Thematizer indeed ensues automatically, manual verification and correction of the output remains with index identification and thematic progression.

Program accuracy thus presents the greatest hurdle to overcome, which invariably affects how users should use and interpret the data output in Thematizer's web interface. False positives and false negatives could lead to confusion on the user's part as seemingly contradictory information could be present in the output. For example, in the index identification task, partial highlights in hypotactic constructions could blur users' understanding of hypotactic themes. Similarly, the incorrect identification of connecting elements in the thematic progression output could lead the user to assume structural or constructional errors in their text. Since there is no way for users to inquire about clarification of discrepancies in their output, this may lead to frustration in Thematizer's use and a questionable reliability of its results. Clarification on how the user should use and interpret the results are provided within each tab, but users may not read them or understand how to apply those explanations to their own text results. On this point specifically, user experience (UX) design surveys would help to isolate usability issues and formulate facilitative instructions, explanations and clarifications.

In spite of these errors, Thematizer represents a marked improvement over previous computational approaches to the automated analysis of thematic structure in terms of its parsing functionality. Compared to the tools developed by Domínguez et al. (2020), Park & Lu (2015) and Schwarz et al. (2008), Thematizer fills a critical gap in its analysis: intertextuality. Instead of uploading texts individually and manually comparing the results, Thematizer allows the user to upload multiple texts and have their analytical results shown directly in the program and in comparative form. As Thematizer was able to analyze 30 training texts averaging 1600 tokens each in less than three minutes, large numbers of texts can be processed in batches and considerably more quickly than manual analyses would allow.[24] Further, since users can

[24] This parsing time is based on the use of a Mac Mini with an M1 processor, MacOS Monterey operating system and 16GB of RAM. Other processors and computer configurations will have varying parsing times required for analysis.

download the entirety of their results, also for multiple texts that were uploaded, this output can be used for tagged corpora construction or as input for research that requires the output data.

The visual breakdown of a text's thematic structure into four separate tabs in Thematizer's web interface illustrates an additional improvement to the tools previously developed. Herein lies Thematizer's strength. Through its fine-grained analyses and corresponding output, Thematizer can visualize the underlying thematic structure evident in the user's text. Access to the discursive and structural development of the user's text can thereby be more readily achieved in an automated and independent manner. It is the individual output produced and presented in the web interface itself that can inform users of their text construction and development via the thematic framework.

Thematizer's first parsing task, index identification, produces a breakdown of the text input in terms of thematic and rhematic constituents in its first tab. Additionally, the distribution of marked themes and their realization within the text is presented graphically above the text. Through a visualization of thematic structure, users may gain initial insights into how they structure their sentences on the whole: Are their themes consistently more complex than their rhemes? Do users repeat the same grammatical themes and in the same form throughout the text? Do they tend to employ multiple marked themes to introduce their sentences or do they hardly employ any at all?

While writers are likely aware of how they construct their texts, either at the sentence level or beyond, they may be unaware of a preponderance of certain realizational patterns. For instance, Thematizer's output could indicate that a user tends to follow up projecting themes with hypotactic themes, thereby creating a complex subordinating structure. Through explicit visualization of such patterns, users can decide to continue with this syntactic patterning or employ alternative realizations for stylistic reasons. Additionally, with the marked theme distribution figure presented at the top of the index identification tab, users may identify patterns of marked theme use, e.g., frequent use of modal themes in their engineering report. When comparing the distribution of marked themes from their own text to other similar text types, they may become aware of their own implicit adherence to or deviation from text type conventions and characteristics.

Identification and subsequent visualization of marked themes, grammatical themes and rhemes thereby affords a degree of dynamicism and discernable feedback on the construction of a user's own text. Particularly for those who pay less attention to how they connect and develop their sentences, this visualization can make overt what users may have originally been unaware of in their writing. With the initial impressions users gained from the analytical results presented in Thematizer's first tab, they are initially exposed to how thematic structure applies to their own writing.

Thematizer's inclusion of the unmarked grammatical theme in the thematic analysis irrespective of the presence of a marked theme represents a theoretical deviation from previous research. Although identification of marked and grammatical themes proved variable in its accuracy, the analysis and visualization of grammatical themes parallel to marked themes represents an improvement to previous thematic theory models. This then fills a gap in previous research, which formed the basis of the present work's first research question. By identifying and tracing grammatical themes throughout the text, users can follow the development of previously established discourse topics that are instantiated as (part of) the foundation of their discourse message. These contribute to a text's method of development, which centers around the experiential topic of a sentence, i.e., the grammatical theme (Fries 1995: 319). As such

topical elements were relegated to rhemes in previous approaches so long as a marked theme was present, the text's (thematic) method of development could become blurred. In Thematizer, however, both marked and unmarked themes were tracked simultaneously, which affords a more comprehensive analysis of discourse topics developed as the foundation of the discourse message. This has the added benefit of tracing GIVEN discourse topics in text analysis from the perspective of information status. Since rhemes are identified and extracted in juxtaposition to marked and unmarked themes, the user can also more readily trace how NEW discourse topics are presented in the rhemes to develop the discourse message.

Thematizer's analysis and classification of marked themes is its second thematic parse, whose results are presented in the second output tab. There, users are presented with the tallied frequencies for their text's marked themes and their semantic classification. Marked theme classification represents a key improvement to previous research, which either ignored marked themes entirely (Domínguez et al. 2020) or subsumed them under a single marked theme class (Park & Lu 2015, Schwarz et al. 2008). The extraction and presentation of the marked themes as used in the text can raise awareness of the types of marked themes the user tends to employ. Not only can users see how the marked theme was realized alone but also how it was used within the sentence from their own text. Additionally, the frequency of these themes can indicate whether users may prioritize certain marked theme types over others. The results here can therefore shed light on the diversity in marked theme use – or even lack thereof. Through multi-document comparisons, users can glean how marked themes are used in similar and divergent ways from their own texts. This can expand their understanding of marked themes' use and realizational patterns.

Aside from the actual use and frequency of marked themes in their texts, the results in this tab can provide insight into the logical and cohesive function that marked themes afford. Particularly for writers at the beginner or intermediate level, exemplification of marked themes in context can reinforce their understanding of marked themes' meaning and use. Additionally, the cohesion that marked themes afford across sentences is presented with respect to the logical development imparted through their semantic classification.

Breaking down marked themes into their semantic subcategories adds another level of detail that the marked theme type, e.g., circumstantial or structural theme, generalizes. For general users, the linguistic terminology used, such as DESIDERATIVE for modal themes or EXPANDING ELABORATIVE for structural themes, may obfuscate the actual semantic contribution a marked theme offers. For linguistics, however, the marked themes' semantic categorization may be of greater use since their automated identification and classification can expedite text analysis. The semantic subcategories in particular can be used to reveal characteristics unique to or shared across text types.

The greatest disadvantage to how the data is presented, both in this tab and others, is its static nature. Users upload their text, whose results are returned as a singular output. Any changes the user may wish to make on the basis of the output would require uploading and analyzing the altered text again. Programming Thematizer such that changes could be dynamically made to the uploaded text would highlight the effect such changes have on the resulting output. That being said, changes can currently be traced by uploading the original and altered versions of their text. Then, the user could compare the original text with those containing different marked themes or progression patterns. In doing so, ad-hoc dynamicism could be achieved, albeit in a somewhat cumbersome manner since changes to the texts would have to occur outside of the program first.

Adding in on-the-fly changes to uploaded texts would have required the use of storing the analytical results in a database. While minimal database creation and queries could have been incorporated, it was decided to first ensure the primarily functionality of Thematizer before data storage and retrieval were implemented. Still, this is a natural progression in the program's development as a means to collect examples of marked themes and their realizations. Instead of depending on uploaded texts alone for comparative purposes, the use of previously stored data for comparison could enrich the analytical output of the user's text. For example, if the user uploaded a text that had a high frequency of the TEMPORAL hypotactic (e.g., *when*) additional examples could be extracted from previous analyses and presented alongside the user's output as suggestions for alternative marked themes of the same class. This would facilitate writers' linguistic expression by being shown additional ways to express the same logic or cohesion with different marked themes. In doing so, directly imbuing feedback on the user's text with the results of previous analyses stored in databases could raise awareness of stylistic variation in thematic selection.

Additionally, while the results are currently presented in tabular form, it may have been helpful to provide an additional figure that summarized average frequencies of semantic subclasses with respect to text type. Then, users would have been able to see the distribution compared to the text they uploaded. As this feature can be readily implemented in the web interface, a future version of Thematizer could be programmed to do so accordingly.

Otherwise, errors that emerged during parsing as outlined in Chapter 6.2 could impair the user's understanding of marked themes and their realization. While the explanation card for the second tab provides general information about the various marked theme types, it may remain unclear why their text has diverse – and even erroneous – realizational patterns. Errors in output in the form of noise or partially extracted marked themes could result in more confusion than clarity about marked themes. The high F_1 scores for marked theme extraction may prevent errors from occurring frequently, but the likelihood of errors from parses persists nonetheless.

Further, appropriate use of marked themes is not considered in the output. Thematizer is only tasked with the extraction of marked themes as used in the uploaded text. Therefore, insufficient coherence on account of infelicitous cohesive devices could be present but remain ignored in the parse. The presentation of information alone without consideration of appropriate usage could lead users to draw false conclusions about marked themes' use. Semantic tests would be required to evaluate how marked themes are used and whether their use would substantiate the logical and structural development across sentences. This, too, represents an additional feature that could be added to Thematizer's functionality for additional, purposeful feedback on the uploaded text.

How a text is developed from sentence to sentence is summarized in the thematic progression classification output in Thematizer's third tab. The two figures that tally the frequency of progression patterns employed and means of progression in the tab may find greater use in linguists' text analyses. This is enriched through Thematizer's ability to compare progression patterns across multiple texts through intertextual analyses within the same figures. Such data can then serve to reinforce or contradict findings on progression pattern frequency in certain text types. Such frequency tallies and visualization of thematic progression using the user's texts are a defining functional improvement over previous computational approaches to thematic analyses.

The output presented in the user's highlighted text with corresponding progression patterns in the third tab affords users many benefits. Firstly, it makes overt how the user developed their

discourse from one sentence to the next. In other words, the output could raise users' awareness of how thematic progression and text structure are closely intertwined. This could encourage reflection of the propositional and structural development of their text through guidance from the highlights. If the output indicated that rhematic progression was present, i.e., the development of rhemes across sentence clusters, the user could consider whether constant or linear progression would be more appropriate given the discourse topic at that point of the text. The output could thereby function as a starting point for the user's reflection on the text's structure and development.

Secondly, frequency tallies could indicate users' preponderance of specific thematic progression patterns. If results indicated that the majority of their text was realized through gapped or constant progression, then users could alter how content is realized thematically or rhematically to diversify thematic progression patterning and ease comprehensibility. Instead of repeated use of continuous progression, which could impart a predictable but potentially monotonous patterning to their text development, conscious revision to more linear progression patterns could diversify their sentence structure. This would impart greater dynamicism to a writer's text while ensuring that logical development by means of thematic progression is ensured.

Thirdly, connecting elements shown in the mouseover results could facilitate confirming or questioning how the discourse topic was progressed across sentence clusters. If Thematizer identified a connecting element two sentences prior, i.e., gapped progression, but the user intended on progression from the preceding sentence, this output could prompt the writer to alter the structure and content of the text at that point. The results could aid in ensuring that developments in the user's text are made concrete, either through lexical repetition, coreference, paraphrase or other cohesive and coherence devices. Logical gaps across sentence clusters could thereby be minimized or ameliorated during the revision process.

Despite the aid these results may provide, the output in this tab proved most problematic on account of the low F_1 scores Thematizer achieved. The overall accuracy Thematizer achieved suggests questionable reliability, particularly on a text-by-text basis. Particularly for linguists who wish to use the data for their own research, the wide range in accuracy would necessitate closer examination of the output. Although the highlighted output with connecting elements and means of thematic progression expediates data validation, ensuring data correctness manually would require added time and effort regardless.

For non-specialists and those less acquainted with thematic progression, there is the risk of users accepting the output at face value. In other words, users may consider the output as true and free of errors. Conflicts between the output and how the author intended on developing the text could thus arise. While users should be critical of the output regardless, drawing false conclusions on the basis of the output for thematic progression could be counterintuitive to the purpose of the program.

It was for this reason specifically that the explanation card stresses the structuring function of thematic progression. Asserting a certain thematic progression frequency distribution for their text (type) would presuppose a prescriptive nature for the program, i.e., stating that the text type users uploaded should or must reflect a certain frequency in thematic patterns would be a fallacy. Although text types have shown to leverage particular thematic patterns over others (cf. Hawes 2010b and 2015), this characteristic alone proved to be insufficient in determining text type membership (cf. Chapter 5.4). Instead, thematic progression frequency should be considered along with other texture characteristics, such as lexical density, ratio of coordinating to

subordinating clauses and means of thematic pattern realization. Stark deviations from typical thematic pattern frequencies may call into question the text's structure vis-à-vis its text type, but such deviations should not be considered infelicitous through Thematizer's output.

In essence, this output is meant to direct users' attention to which sentence constituents comprise the foundation (theme) or core (rheme) of the message, and how thematic progression reveals their development through textual discourse. If the parser's output facilitates reflected consideration of the user's structural and logical development in the text, then it can be seen as a success. Accurate results for substantiation of the thematic progression patterns present would reinforce the actual (the user's) versus the assumed (the parser's) text structure. This then motivates further work on Thematizer such that harmony between the actual and assumed text structure can be achieved and reflected in the output.

A final note concerns the design and presentation of data for the third tab in particular. The current version enables quick access to the relevant thematic progression data through the figures, highlights and mouseover data. This is a marked improvement over the original design, where all thematic patterns were presented in a table alone. Identifying the connecting elements and development of discourse topics, let alone structure across sentence clusters, was unintuitive and cumbersome. For that reason, the same scheme used for indicating theme and rheme spans in the first tab was adopted and expanded for thematic progression classification. This first facilitates a generalized summary of the thematic patterns and means of connection through the frequency charts. Secondly, the text can be perused in its entirety with relevant information about how the sentence was developed thematically or rhematically as well as by which realizational means. Users can thereby closely track discourse development from one sentence to the next.

Where the output presentation lacks is discourse development at the macro level. A text's macrothemes may provide insight into dominant, overarching discourse topics, but the topics derived from macrothemes and within the individual rhetorical sections remain harder to track. With the highlighted output, users can progress from the themes and rhemes in one sentence to the next to trace discourse development. Yet, this requires examination of each sentence individually. In addition to data presented for discourse development at the micro (i.e., sentence) level, presentation of how the themes and rhemes contributed to discourse development at the macro level would enrich the analysis. This could be summarized as a list of themes and rhemes for quick reference. Otherwise, paraphrasing the discourse topics on a paragraph-by-paragraph basis would achieve a greater reflection of the macro-level discourse topics. The important point here would be to employ the discourse topics only from the themes and rhemes that contributed to thematic progression. Doing so would reduce content noise in the paraphrase while ensuring that the relevant discourse topics involved in the text's thematic progression are covered. With such output, a breakdown of the text's development in terms of discourse topics would introduce the results for thematic progression. Then, the minutiae of the progression patterns and means of pattern realization could be examined with the subsequent highlighted output.

In the fourth and final results tab, users are able to examine how their text compares to the five text types used for the development of Thematizer: Wikipedia articles, blog articles, lyrics, L1 university texts and L2 university texts. Frequencies of marked themes, thematic progression patterns and means of realization are summarized in their respective graph and used as data points for comparison against the user's text. The purpose behind this tab was, again, a descriptive one. The graphs provide a generalized indication of frequency distributions in the given text type. These, in turn, indicate where a user's text falls within these frequency distributions and with respect to text type.

The output here can thus indicate whether their text adheres to or deviates from marked theme and thematic progression usage. With this data, users can then make adjustments to their text in order for it to be more in line with a text type's average frequencies. Conversely, the user can consciously choose to ignore these frequencies to make their text stand out structurally and stylistically from others within the same text type. Again, the purpose behind these figures and data was not to show users the frequency distributions their text should or must have. Instead, they were meant to bring to light the linguistic and structural choices they made, consciously or not, in the development of their text. As structure and thematic selection may not be factors that writers consciously consider during the writing process, such figures can illuminate underlying stylistic and idiosyncratic tendencies that users may have otherwise remained unaware of.

Ultimately, these results provide insight into the intertextuality between the user's text and those of (dis)similar text types. For linguists, this data can be used for confirmation or repudiation of linguistic phenomena in text. For example, the lower number of projecting themes in informal text types proved true in the training texts analyzed: these occurred at a frequency of 5.7% in informal texts compared to 14.6% in formal texts. As projecting themes are commonly used to make formulations objective, e.g., *it could be argued that […]* instead of *I think that […]*, their frequent use can be understood as a reflection of the conventions for texts of formal register. Further, their infrequent use in less formal texts reinforces this finding. A similarly high frequency of both lexical repetition and paraphrase could be indicative of a formal-register text, whereas frequent use of one alone may be more emblematic of less formal text. Through cumulative analysis, the degree to which each of these factors pervade the text type can be readily investigated with the output in the final tab of Thematizer.

For non-native writers, particularly at the beginner levels, the results can also be a boon. Knowing whether their text is close to or far from the standard can foster written expression that is unfamiliar to them or underdeveloped. Instead of relying on lexical repetition alone, employing paraphrase and coreference more frequently could afford greater structural and stylistic diversity to their text. The incorporation of fewer but a wider range of marked themes, as native speakers tend to do (cf. Hinkel 2001), could allow their texts to achieve greater idiomaticity. Finally, a deeper awareness of sentence structure and development in the case of English specifically could be another consequence of the thematic output.

In spite of the potential use cases this output enables, limitations to its applicability are present here as well. Firstly, the frequency results for each text type are minimally representative on account of the small number of test texts employed. Only 30 texts per text type were used to produce the average frequency figures, which represents a mere subsampling of the text type. Therefore, in its current form, representativity cannot be justifiably claimed. This has been noted in the explanation card but may remain ignored and overlooked during perusal of the results. Considerably more texts of each text type would be required to achieve greater representativity for the generalized findings in each figure. Once prominent parsing errors have been addressed in future versions of Thematizer, the number of text types and their frequencies could be expanded and reflected here.

Next, the interpretability of the results here could potentially come at the expense of user's creativity or idiosyncratic writing style. A user's text may differ considerably from the frequency tallies in each figure. This may prompt the user to reformulate and restructure their text so that it might adhere to the average more closely. While such actions are an additional linguistic and stylistic choice, they could stifle the creativity that the original text once had. Exceptionalism through written expression could be undermined as a result of the user desiring

to fall into the average through changes the results may encourage. Both novelty in expression and the writer's voice could therefore be dampened if the results were taken as a 'should-be' instead of a 'can-be' case.

Finally, as is possible in the previous tabs as well, the user simply might not understand the relevance of the data that the three figures in tab four summarize. They may see that the frequency distributions in their text differ or equate to those of other text types. However, what greater meaning the results have could remain nebulous. Also, if a user's text did not fall within the average frequency of the same text type, they may interpret the result as 'bad' and, by extension, their text as poor writing. The results could thereby discourage learners of English in their writing skills and endeavors.

In summary, the functionality and results that Thematizer provide appear to be both bound by and independent of the automated, unsupervised feedback it facilitates. Underlying structural characteristics and linguistic choices expressed through marked themes and thematic progression are visualized through the summarized results and the accompanying highlights of the user's text. How users can draw conclusions from their text's results, while explained in each tab's cards, might not ensure comprehension of the output or the intended purpose. A prescriptive, as opposed to a descriptive, assumption of the tool's output on the user's end could further hamper the interpretability of the output. Finally, recurring errors that pervade the text's analysis reduce the reliability and comprehensibility of the automated output. In-person guidance of the results would clarify the understanding of the results and purpose thereof but thereby counteract the self-driven usage of the tool itself.

The drawbacks that persist in the current form of Thematizer cannot be ignored. Yet, it is the insights that the software provides that encourage continued development of the program. Accurate and fine-tuned feedback on a text's method of development as a reflection of linguistic choice is a complex undertaking that Thematizer attempts to present in a simple and intuitive manner. Just as other online writing tools aim to enrich the writing process and outcome, input from Thematizer helps bring to light qualities of a text that the user may have remained unaware of during the writing process, such as the reversal of GIVEN and NEW information or the propensity towards a single marked theme class as a cohesive device. In doing so, Thematizer facilitates greater access to and understanding of thematic theory both in their own and others' writing. While future work will first and foremost aim to improve parsing accuracy for operationalization of thematic theory, in its current form, Thematizer fills a marked gap in how thematic structure is automatically analyzed and made accessible to writers through computational and visualization means.

Chapter 7 – Conclusion

The purpose of this final chapter is to summarize the present work on thematic theory and its findings. To that end, answers to the two core research questions as conclusions drawn from the findings begin the discussion. This will aid in contextualizing the key findings that came to define the work. On the basis of these key findings and conclusions, contributions to thematic theory and limitations of the current work are subsequently outlined. The chapter then concludes with potential research opportunities that the present work enables.

Investigating thematic theory from both a theoretical and computational perspective formed the foundation of the present dissertation. Theoretical models informed by the historical development of thematic theory from the Prague School of Linguistics and systemic functional grammar served as the framework for the development of Thematizer, which constituted the final product of this dissertation. This software represents the application of thematic theory via computational means by automating the analysis of text in terms of themes, rhemes and thematic progression.

Both the theoretical and computational approach to thematic theory in the present work was motivated by two core research questions: Firstly, what deficiencies exist in contemporary thematic models and how can they be overcome? Secondly, how can thematic theory be operationalized by computational means so as to make it accessible to writers?

Starting with deficiencies identified in previous research, Thematizer deviates from the Hallidayan approach in how unmarked themes are analyzed in conjunction with marked themes. Whereas the presence of a marked theme in the Hallidayan framework forced the unmarked, topical theme to be subsumed under the rheme, the present work allows both marked and unmarked themes to constitute the overall theme if realized together. The justification behind this analytical approach was due to the discursive function that themes fulfill. These form the foundation of the discourse message on account of their GIVEN information status upon being explicitly realized previously in the text. Therefore, if they were relegated to the rheme instead, which is where NEW discourse topics are introduced, then their contribution to discourse development would be conflated and obscured by its rhematic membership. By separating marked themes from the unmarked theme (denoted as the grammatical theme in the present work), GIVEN discourse topics could be more readily traced as the text's method of development and foundation of the discourse message. In turn, their foundational function and thematic realization could then be juxtaposed with the rhematic discourse topics, whose purpose is the furthering of the discourse through NEW topics. This facilitated a clearer distinction between thematic and rhematic elements and how they contribute to the unfolding discourse.

The second gap identified in previous research was the treatment of marked themes. In their analysis, Thematizer delineates marked themes into structural, hypotactic, projecting, circumstantial and modal themes. This is an expansion to how marked themes are analyzed, which historically were first subsumed under the general metafunction of interpersonal or textual themes. The present work's dissection of marked themes into their respective types allowed for their functional category to be associated with their textual and discursive realization. For instance, identifying a marked theme as a structural theme indicates its functional class of contributing to the logical structure of the text, which highlights its nature of establishing cohesion and coherence between sentences. Categorizing marked themes into one of the five classes thereby allowed for a fine-grained analysis of how they contributed to the contextualization of the theme and rheme that followed and to the development of discourse at that point in the text.

Further, while previous text linguistics work has categorized marked themes into their semantic subclasses (cf. Halliday & Matthiessen 2014: 107), this step has yet to be automated by computational means in previous software tools. Upon classification of the marked theme, Thematizer identifies the marked theme's respective semantic subclass for even greater insight into its semantic contribution to the text. For example, the circumstantial marked theme adjunct *in* can indicate TEMPORALITY, LOCATION, MANNER and more, depending on how it is realized syntactically in the text. As semantic classification has typically been conducted manually in previous text linguistics studies, Thematizer's ability to automatically and accurately classify marked themes' semantic subclass can quickly facilitate the identification of their discursive function and semantics in text without manual analysis.

Next, the poor visualization of thematic constituents and thematic progression in previous software tools was identified as a key drawback to how they made thematic theory accessible to users. Previous work visualized thematic progression within an Excel table (cf. Leong 2019), node maps (cf. Domínguez et al. 2020) or textually (Schwarz et al. 2008; Park & Lu 2015). While such visualization captured thematic elements and progression, they were extracted from the text in which they were used. Thus, users were required to refer to the original text for comparison against the thematic analysis output. Thematizer overcame this functional and presentational deficiency by embedding the analytical results within the user's text through visual highlights and mouseover explanations. Users can thereby identify which sentence constituents are thematic or rhematic, which marked theme class they belong to, which thematic progression patterns are present and how these are realized in the user's text all at once. This improves both usability and comprehensibility of analytical output and, in turn, accessibility to thematic theory.

Yet another deficiency in previous research on thematic theory and its automated analysis was found in the number of documents that software could analyze. Compared to previous tools, Thematizer is able to process one or more texts simultaneously, which offers greater analytical functionality and output while facilitating intertextual analyses. If the user uploads more than one text for analysis, they can compare the frequency distributions of the thematic constituents, thematic progression patterns and means of progression among each text at the same time. This obviates the need for analyzing texts individually, collating the results on one's own and then manually comparing the results against each other. Thematizer's multi-document analyses thereby facilitate intertextual insights with respect to thematic theory, which can illuminate texture characteristics across (dis)similar text types.

The final gap in previous research was a lack of access to the results from the thematic analyses. Upon completion of the thematic analysis, users have the option of saving the results as a CSV, Excel or JSON file. The analytical output not only includes dependency parses but also thematic and rhematic constituents, marked theme analysis and classification, thematic progression classification and means of thematic progression. Users can use this data to facilitate and expedite the analysis of texts for annotated corpora or text linguistics studies. As no publicly available corpora that have been annotated in terms of thematic constituency exist at the time of this writing, the automated output that Thematizer produces can aid in the creation of such corpora.

Against the backdrop of these theoretical and functional deficiencies, Thematizer was developed to operationalize thematic theory by computational means and make thematic theory accessible to writers. Thematizer's degree of success was measured in terms of its parsing accuracy for each of its three thematic parses, as captured by the F_1 score. In its first parsing task, Thematizer achieved $F_1 = 85.8\%$ for the training dataset and $F_1 = 92.0\%$ for the validation

(test) dataset when identifying marked themes, grammatical themes and rhemes. Since not all texts were able to unilaterally achieve the gold standard of 89.1%, Thematizer's operationalization of thematic theory in terms of index identification was deemed partially successful.

The variable accuracy rates for the identification of themes and rhemes was shown to stem from the dependency parses that index identification relied on. So long as the dependency parse returned was correct, Thematizer was able to isolate thematic and rhematic constituents correctly. Where analysis encountered difficulty, however, was with misparsed t-units, ungrammatical text, syntactic ambiguity and clerical errors. It was concluded that complementing dependency parses with part-of-speech tagging and pattern matching as a biconditional approach to index identification could improve the overall accuracy of Thematizer's first parsing task.

While breaking down sentences into smaller thematic and rhematic constituents facilitated a fine-grained analysis of their contribution to discourse development, the addition of a grammatical theme compounded parsing issues that may have been absent without its added treatment. This was further complicated by the myriad ways in which thematic and rhematic constituents can be syntactically realized. That being said, an average F_1 score of 88.9% between both datasets indicate a sound theoretical and programmatic approach to index identification.

Thematizer's second thematic parse, marked theme identification and classification, proved to be the most accurate of all three parses. It thereby represents the sole case of successful operationalization of thematic theory in terms of marked themes. For the training texts, Thematizer yielded an F_1 score of 94.9%, which was slightly higher than that of the test dataset at 93.4%. On account of the high accuracy rates, Thematizer's output for marked theme classification can be seen as reliable and facilitative of accessibility to thematic theory. This was achieved through the biconditional testing methodology of dependency and part-of-speech parsing and pattern matching. As such biconditional testing parameters proved successful in this task, it reinforces the finding that such an approach could be implemented in the index identification task to achieve greater accuracy rates.

Marked theme parsing was most readily impaired by the resolution of right edge dependents, which were used to extract marked theme spans in their entirety. Erroneous resolution ultimately led to the partial extraction of marked theme spans, which occasionally affected their semantic subclassification. Additionally, parsing errors that emerged in the index identification tasks manifested in this second parse as well, hampering the corresponding accuracy occasionally but minimally.

The final finding from marked theme classification was the relationship between their frequency distribution and text register. Results indicated that syntactically and semantically more complex themes (circumstantial, projecting and, partially, hypotactic themes) were more frequent in formal-register texts. Conversely, comparatively simpler themes (structural and modal themes) were found to be representative of texts with a less formal register. Therefore, a (lack of) complexity in both subject matter and formality of text type showed to be reflected in the usage of particular marked themes.

Thematizer's third and final thematic parse, thematic progression classification, achieved the lowest F_1 scores and exhibited the widest range of parsing accuracy of all three parsing tasks. Specifically, training texts were more accurate with $F_1 = 80.2\%$, while the test dataset ultimately

yielded $F_1 = 75.9\%$. As neither the training nor the test dataset was able to consistently exceed the gold standard of 79.2% in the individual text analyses, Thematizer's operationalization of thematic theory in terms of thematic progression was concluded to be incomplete.

The reason for this was attributed to three primary faults in the parsing and programming methodology: the use of cosine similarity with pre-defined upper and lower bounds, erroneous coreference resolution, and the order of the thematic progression tests. Cosine similarity tests were employed to resolve lexical entailment, which includes but is not limited to hypernymy/hyponymy, meronymy and ellipsis. The values returned by these tests were then used to determine the corresponding thematic progression pattern across sentence clusters. However, the pre-defined upper and lower bounds for the similarity values proved too restrictive. This caused gapped progression patterns to become overinflated in their identification as well as constant continuous and linear progression to be frequently overlooked. It was therefore concluded that an alternative approach to upper and lower bounds should be pursued in order to distinguish between the various thematic progression patterns via lexical entailment.

Secondly, Thematizer experienced numerous parsing errors when resolving coreference by coreference chains and their corresponding indices in the text. While faulty coreference chains, i.e., false proform-antecedent pairs, constituted one group of errors, Thematizer's use of the textual realization of antecedents instead of their indices constituted the other. It was concluded that the use of coreference chain indices, which already indicate the location of the antecedent to resolve in the text, would not only be programmatically simpler but also more accurate in their parse.

Finally, the order in which the various tests were conducted to determine the thematic progression pattern evident across sentence clusters was also scrutinized. In its current version, Thematizer tests for coreference resolution first, followed by lexical repetition, macrotheme instantiation, lexical entailment and finally thematic breaks. As lexical repetition proved to be the most common means of progression in the datasets, this should have been carried out first in order to account for its prevalence in thematic progression. A parametrically weighted approach to thematic progression classification was also suggested as a possible means for assigning priority to particular patterns if multiple means of progression were found to exist across sentence clusters.

Residual errors to have affected Thematizer's thematic progression parse were identified in faulty lexical repetition tests and an overgeneralization of thematic breaks on occasion. Findings indicated that a bidirectional search pattern for lexical repetition could more readily account for the repetition of lexemes realized as varying parts of speech across sentence clusters. Additionally, it was argued that removing thematic breaks as a default fail case after Thematizer has progressed through the previous thematic progression tests could reduce the inflated number of thematic breaks identified. Instead, the use of cohesion metrics could aid in substantiating the presence of a thematic break as an indication of a rhetorical shift to a new section in the text.

The final finding to emerge from the thematic progression analyses was the relationship between the frequency of thematic progression patterns and text type. It was found that thematic progression patterns alone cannot be used to define a text type's membership but instead should be considered an additional texture characteristic. This was due to the numerous pairs of thematic progression patterns that shared a statistically significant relationship with multiple text types. The sole exception was that of gapped linear progression with L1 university texts,

which should be scrutinized nonetheless on account of linear progression's failure to achieve statistical significance in L1 university texts.

In light of these core findings, the contribution that Thematizer affords to an understanding of and accessibility to thematic theory can be reviewed. Aside from the contributions outlined above by overcoming deficiencies from previous research, the data visualization and analytical summaries that Thematizer provides enable the thematic structure underlying a user's text to be brought to the fore. Through highlights and generalized explanations of the analytical output, users can identify: the constituents responsible for establishing the foundation of their discourse message (grammatical themes); for contextualizing and framing the discourse message (marked themes); and for developing the discourse message further (rhemes). Through thematic progression, users can trace how discourse topics and marked themes are developed throughout a text and compare these findings to other text types.

While this does not necessarily guarantee an understanding of thematic theory, the output that Thematizer produces can illuminate the structural composition of a user's text in terms of thematic constituents. In doing so, a keener reflection on how the user consciously selects previously established and new topics to develop the discourse may be facilitated. The output can also shed light on structural, compositional or stylistic tendencies that the user may have been unaware of in their own writing. This can then foster further reflection on how users may compose their texts.

As Thematizer is intended as an automated tool for feedback on a user's writing, teachers can employ Thematizer for didactic purposes. Particularly for non-native English users of all levels, Thematizer can be an addition to their writing toolkit. As previous research has shown the benefits of applying the thematic paradigm to students' writing and understanding of language-specific compositional structuring, this tool can aid in the continued improvement of students' writing skills (McCabe & Belmonte 1998; Green et al. 2000; Gunawan & Aziza 2017; Naderi & Koohestanian 2014). Teachers can also make use of the tool to provide immediate and visualized feedback on their students' texts with a particular focus on sentence construction, text structure and text development. Together with other texture characteristics, such as lexical density, readability scores, cohesion and coherence, Thematizer can provide concrete input on students' writing.

Finally, as mentioned previously, researchers can also benefit from the automated thematic analyses that Thematizer enables. Instead of manually dissecting texts in terms of their thematic structure and development, Thematizer can expedite the process while providing visualized output for confirmation of the results. Since the output can then be exported, the analytical data can be used to complement or enrich text linguistics studies. This provides agency over linguists' own text analyses and allows leveraging the output for their particular research.

Despite the theoretical and practical contributions Thematizer has enabled, both the present research and the developed software were subject to a number of limitations. First and foremost, the variable accuracy rates achieved in the index identification and thematic progression tasks merit questioning the reliability of the output. While some texts were able to exceed the gold standard in all three parsing tasks, these were the exception to the rule. This is problematic for both linguists who wish to use the output for their own analyses and for writers who wish to gain feedback on their text development. Errors in the thematic analyses could cause users to draw incorrect conclusions about how their text is structured. In turn, users may apply those false conclusions to future compositions. Although they are encouraged to question Thematizer's output as a way to reflect on the structure and development of their text, this

cannot be ensured due to the unsupervised nature of Thematizer's feedback. The variable accuracy rates have been mentioned in the web interface itself, but, again, acknowledging this requires the user to read through the explanations provided in the web interface. Resolving the parsing issues outlined in Chapter 6 to increase the overall parsing accuracy therefore forms the primary impetus behind future work on Thematizer.

Researcher bias forms another limitation of the present study. All training and case study texts were analyzed by the present author alone. Corroboration of the thematic progression patterns and marked theme subclasses in the training and test datasets is therefore absent. Further, clerical and analytical errors that could have been avoided through corroboration may have also affected the parser's functionality and output. In order to validate the present author's thematic analyses, other researchers would need to analyze the training and test texts for comparison. A t-test could then determine whether the analytical results differ from one another. Should no significant difference emerge between the analyzed output, then this would lend credence to the validity of the present author's analyses.

An implicit bias may have also been introduced in how thematic progression patterns are classified. Since testing is done linearly, greater precedence is implicitly given to tests that take place first. Although lexical repetition was found to be the most frequent means of progression (49.8%), testing for it occurred after coreference resolution. Otherwise, coreference resolution might have been largely overlooked due to the prevalence of lexical repetition. As mentioned above, a parametrically weighted approach could aid in reducing progression pattern bias.

Sampling issues and sample size of the texts chosen for training and validation constitute the next limitation of the present work. The present study exceeded the number of texts most previous research used in thematic progression analysis by 50–140 texts, which enabled more extensive text analyses. However, limiting the training texts to five text types with 30 text samples each is hardly sufficient for capturing the entirety of their texture characteristics. A greater number of texts would be required for more robust statistical testing and in order for Thematizer to capture additional texture characteristics emblematic of a specific text type.

Sample size constraints are particularly evident in the number of test texts employed for validating Thematizer's functionality and output. These amounted to ten test texts in total due to time constraints specifically. The manual analysis of the test texts and manual correction of errors from the output for data analysis prevented further addition of validation texts. As a result, the training-test ratio of all texts considered was 90.0% : 10.0%. Generally, an 80.0% : 20.0% ratio is considered the standard training-test ratio in computational analyses (Gholamy et al. 2018). Therefore, an additional 10.0% of test texts would have been required to reach this ratio. In doing so, further texts for data validation could have provided greater evidence for or against the parsing accuracy of Thematizer.

The requirement for texts to be in text file format (.txt) for analysis could also prove a hurdle for users. PDF file formats and those produced by common text editors nowadays (Word, Google Docs, Open Office, Pages) are not accepted in Thematizer's parsing functionality due to the formatting applied to such file formats. A text file format requirement therefore necessitates the conversion of other file formats for use and analysis. Texts can simply be copy-pasted into the text field provided in the user interface; however, this is only possible for a single text. Thus, users would need to either save their texts as a text file or convert their files beforehand for multidocument analyses. This represents an additional step before processing that could discourage the use of Thematizer.

Despite these limitations, the development of Thematizer and its underlying theoretical work provide new avenues for research within the linguistics community. Improving the tool's parsing accuracy, particularly for thematic progression classification and index identification, is fundamental to increasing the reliability of the output. The first step is a re-examination of how t-units are parsed since their misparses emerged in every parsing task and introduced a variety of related errors that could have otherwise been mitigated. The next step is to address lexical entailment through the removal or adjustment of upper and lower bounds for cosine similarity values. Removing cosine similarity tests from gapped progression patterns such that lexical repetition alone is considered as means of progression represents an additional potential step in the tool's next developmental version. Otherwise, fixing indexical issues Thematizer had with coreference resolution from conflating clefts with the dummy-*it* in its search pattern methodology would further aid in improving parsing accuracy in index identification and thematic progression classification.

Once accuracy rates reach or exceed the gold standard from previous research, the resulting output can be used more reliably and without manual correction. One use case for future research with the help of Thematizer is the collation of an annotated corpus for thematic progression. To date, no publicly available corpora exist that are labeled in terms of marked themes, their semantic subclasses and corresponding progression patterns. As a significant number of texts are required to draw conclusions about texture characteristics of a given text type, such a corpus would aid in the thematic characterization of texts.

Such a collection of labeled texts would also facilitate training machine learning models for the automated analysis of texts using neural networks or deep learning models. In fact, the original idea behind Thematizer was a thematic progression pipeline that could be incorporated into a natural language processing tool. It was quickly discovered, however, that the requisite labeled texts for model training were unavailable. Labeling would have had to be done manually on thousands of texts to achieve representative and statistically accurate language modeling. Thematizer was therefore conceptualized as a steppingstone to text labeling and annotation.

Through the present work, Thematizer could be used to create the annotated corpus which would become training data for a thematic progression pipeline. It is planned to construct this pipeline as a comparative version for Thematizer: once complete, accuracy rates for the current version of Thematizer can be compared against the future version with deep learning models. As natural language processing applications are rapidly employing approaches based on machine learning, the transformation of Thematizer into a state-of-the-art program would reflect advancements already being made in the computational linguistics community. The use of transformers in the thematic model to identify and track thematic progression in particular represents one way to bring Thematizer's programmatic functionality to a state-of-the-art level (cf. Wolf et al. 2020). A thematic progression pipeline would also be the first of its kind as no such model exists at the time of this writing.

In its current version, Thematizer can be used to automatically analyze texts in terms of their marked themes due to its overall accuracy of 93.4% or higher. Thus, while errors still persist, the few cases that emerge minimally detract from the reliability of the parse data. Together with the extraction of marked themes for frequency distributions, automated semantic subclassification can shed light on the semantic contribution fronted marked adjuncts provide in text. Future work could therefore take the form of intertextual comparative analyses with respect to marked theme use and frequency. Examining marked themes' functions as cohesive ties or contextualizing phrases upon their automatic extraction from text could then inform research on natives' and non-natives' text.

The output that Thematizer produces could also be enhanced by linking thematic progression patterns with rhetorical structures. As Cloran proposed, so-called rhetorical activities are generalized based on the propositional content of each sentence (1995: 362; 364-365), which is then reflected in the content of themes and rhemes. Activities such as account, observation, generalization and commentary could then be derived and compared with the thematic progression patterns in which they emerged. The potential for mapping rhetorical structure at the micro and macro level onto thematic progression patterns by means of Thematizer could then be enabled. Similarly, tracing a text's method of development via thematic progression patterns, as outlined in Fries (1995), could be done more rapidly and with more text samples by using Thematizer. Determining the relative simplicity or complexity of themes' and rhemes' lexical or semantic realization could be examined as a reflection of a text's method of development. How texts are received or comprehended on the basis of thematic complexity could further inform the feedback Thematizer provides to its users. For authors of all levels and language backgrounds, this insight could aid in the production of comprehensible and cogent pieces of writing.

Future research efforts should also examine the often neglected rheme in more detail. Since most research is primarily devoted to the theme and its development in terms of content and discourse, the discourse contribution of the rheme has so far remained rather secondary. Contextualizing the discourse contribution of the rheme almost exclusively in terms of subsequent themes underscores its rather secondary treatment in thematic theory. A more detailed investigation into the development of rhemes' propositional content, as done with thematic progression, could provide further theories on the structural, semantic, and discursive properties of the rheme.

Finally, while the present work focused on the written mode entirely, the analysis of spoken text in terms of its thematic constituents and progression could be facilitated with Thematizer. A difference in speech mode suggests a difference in thematic progression patterns, as research has already shown (Plum 1988; Muttaqin 2017, Smolka 2017). With the ubiquity and increasing accuracy of automated speech-to-text tools, these could be used in conjunction with Thematizer to automate the analysis of thematic progression. Prosodic characteristics could also be examined to investigate their effect on thematic progression realization. In the same vein, discourse situations could be typified similar to text types with regard to which thematic progression patterns are used most frequently and which discourse factors contribute to specific pattern use.

The development of Thematizer as outlined in the present work represents foundational work for bringing the analysis of text, sentence structure and discourse development into the modern age. The tool presents a novel approach to computationally automating the analysis of a text's structural and logical development with respect to thematic theory. It also implements a contemporary, programmatic approach to natural language processing through its multivariate analysis and output. Users can gain an understanding of how they construct their sentences with marked themes and how their discourse is developed over the course of their text through thematic progression. Ultimately, Thematizer brings facets of the user's written language to light through the tool's automated, descriptive and summative feedback as a means to help shape their voice and idiosyncratic expression.

Appendix A

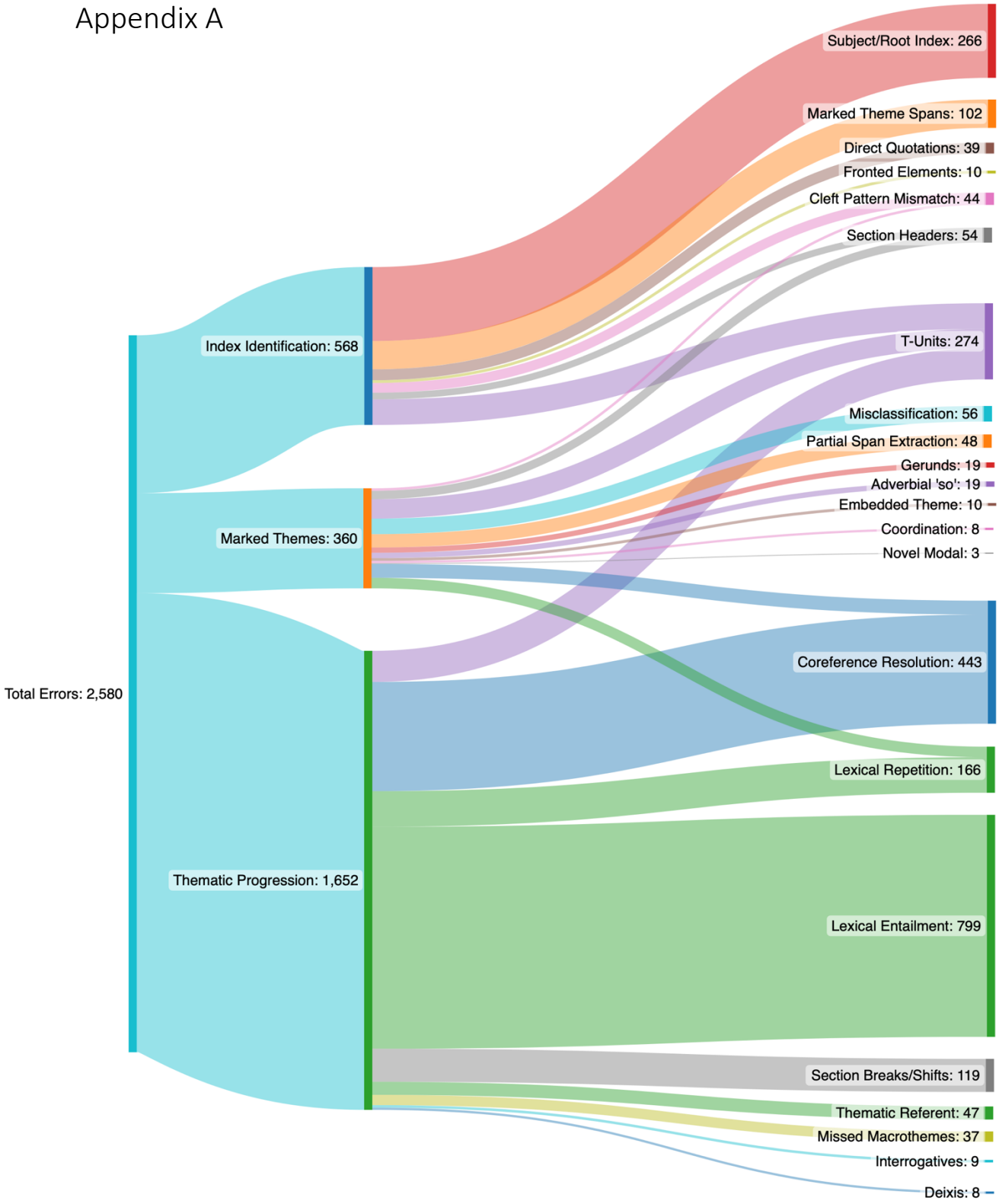


*Figure A-1: Breakdown of each error class for the **training** dataset. Error frequencies are given as total absolute values and split into the three thematic parsing tasks. Errors that affected multiple parsing tasks are indicated by the shared flows, for instance lexical repetition.*

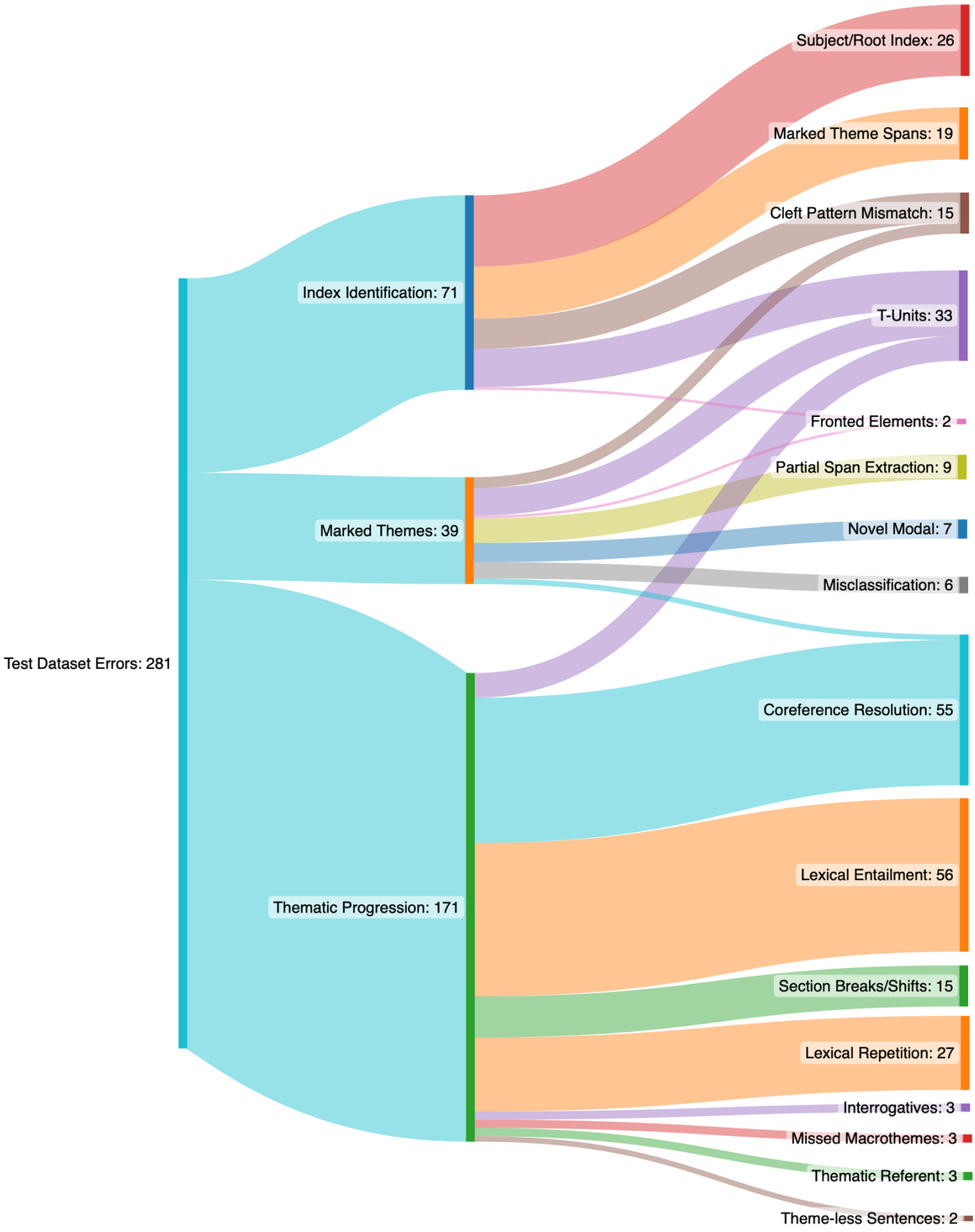


*Figure A-2: Breakdown of each error class for the **test** dataset. Error frequencies are given as total absolute values and split into the three thematic parsing tasks. Errors that affected multiple parsing tasks are indicated by the shared flows, for example t-units.*

		Circumstantial Themes		*Hypotactic Themes*		*Projecting Themes*		*Structural Themes*		*Modal Themes*	
Dataset	Text Type	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Training Texts	Wikipedia Articles	95.1%	89.9%	93.4%	89.1%	95.9%	81.9%	99.3%	94.8%	93.4%	89.1%
	L1 University Texts	93.3%	92.3%	94.2%	96.8%	94.2%	94.2%	95.5%	98.3%	94.2%	96.8%
	Blog Articles	92.7%	92.1%	94.6%	92.6%	95.5%	89.4%	99.4%	97.6%	94.6%	92.6%
	Lyrics	89.8%	88.3%	92.2%	90.4%	95.0%	90.5%	99.0%	96.5%	92.2%	90.4%
	L2 University Texts	95.2%	96.2%	95.9%	95.9%	95.7%	91.8%	99.0%	99.0%	95.9%	95.9%
Test Texts	Gaming News Site	100.0%	100.0%	100.0%	100.0%	90.0%	100.0%	100.0%	95.5%	100.0%	100.0%
	Newspaper Article	90.9%	83.3%	100.0%	100.0%	75.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Linguistics Textbook	83.3%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	80.0%	100.0%	100.0%
	Reddit Comments	66.7%	100.0%	88.9%	100.0%	100.0%	100.0%	100.0%	100.0%	80.0%	50.0%
	Editorial	100.0%	100.0%	100.0%	80.0%	100.0%	100.0%	100.0%	81.3%	100.0%	100.0%
	Obituary	100.0%	85.7%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	100.0%	100.0%
	Blog Comments Section	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	84.6%	100.0%	60.0%
	Wikipedia Article	94.4%	100.0%	83.3%	83.3%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	L1 University Text	100.0%	100.0%	86.7%	100.0%	100.0%	62.5%	100.0%	100.0%	100.0%	100.0%
	Short Story	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	88.9%	100.0%	100.0%	66.7%

Legend	50.0%	60.0%	70.0%	80.0%	90.0%	100.0%

Table A-1: Precision and recall scores for each marked theme class with respect to text type. This breakdown illustrates the degree to which parsing errors permeated marked theme classification in terms of correct classification (precision) and extraction of marked themes present in the texts (recall). Higher precision and recall scores (highlighted in green) indicate a higher overall parsing accuracy.

Precision and Recall Scores of Thematic Progression Patterns from Training Texts								
Text Type	Accuracy Parameter	Constant	Gapped Constant	Gapped Linear	Linear	Macrotheme	Rhematic	Thematic Break
Wikipedia Articles	Precision	94.2%	61.1%	43.8%	64.7%	98.1%	56.6%	43.1%
	Recall	64.2%	86.3%	88.2%	77.1%	87.3%	55.6%	82.0%
Lyrics	Precision	98.0%	94.0%	71.4%	80.1%	100.0%	97.6%	82.1%
	Recall	83.9%	88.7%	94.3%	78.7%	96.0%	90.9%	92.9%
L1 University Texts	Precision	91.8%	71.0%	55.6%	84.8%	96.9%	56.9%	60.4%
	Recall	76.2%	90.8%	95.8%	74.6%	87.0%	72.5%	76.4%
Blog Articles	Precision	89.3%	84.5%	67.9%	81.4%	98.3%	72.6%	86.9%
	Recall	82.2%	86.3%	89.8%	80.1%	96.6%	82.1%	83.2%
L2 University Texts	Precision	92.7%	82.0%	66.3%	85.8%	97.1%	93.6%	69.2%
	Recall	86.1%	90.9%	96.5%	84.2%	89.5%	93.6%	58.1%

Legend	40.0%	55.0%	70.0%	85.0%	100.0%

*Table A-2: Precision and recall scores for each thematic progression pattern from the **training** dataset. This breakdown illustrates the degree to which parsing errors permeated thematic progression classification in terms of correct classification (precision) and extraction of thematic progression patterns present in the texts (recall). Higher precision and recall scores (highlighted in green) indicate a higher overall parsing accuracy.*

Precision and Recall Scores of Thematic Progression Patterns from Test Texts								
Text Type	Accuracy Parameter	Constant	Gapped Constant	Gapped Linear	Linear	Macrotheme	Rhematic	Thematic Break
Gaming News Article	Precision	91.7%	83.3%	75.0%	100.0%	100.0%	100.0%	75.0%
	Recall	91.7%	100.0%	100.0%	82.1%	75.0%	100.0%	100.0%
Newspaper Article	Precision	93.3%	80.0%	64.7%	71.9%	100.0%	100.0%	81.3%
	Recall	68.3%	88.9%	91.7%	82.1%	100.0%	0.0%	100.0%
Linguistics Textbook	Precision	100.0%	80.0%	66.7%	83.3%	100.0%	75.0%	62.5%
	Recall	66.7%	100.0%	100.0%	75.0%	100.0%	100.0%	83.3%
Reddit Comments	Precision	88.9%	100.0%	83.3%	81.0%	100.0%	77.8%	90.0%
	Recall	72.7%	100.0%	100.0%	81.0%	100.0%	93.3%	84.4%
Editorial	Precision	85.0%	72.7%	66.7%	76.5%	100.0%	100.0%	75.0%
	Recall	70.8%	88.9%	85.7%	76.5%	100.0%	100.0%	75.0%
Obituary	Precision	92.9%	70.0%	60.0%	73.3%	66.7%	50.0%	60.0%
	Recall	52.0%	100.0%	100.0%	73.3%	100.0%	100.0%	100.0%
Blog Comments	Precision	93.1%	100.0%	63.6%	81.8%	100.0%	100.0%	78.6%
	Recall	90.0%	83.3%	77.8%	81.8%	100.0%	60.0%	100.0%
Wikipedia Article	Precision	81.8%	62.5%	56.3%	78.4%	100.0%	100.0%	80.0%
	Recall	39.1%	55.6%	100.0%	90.6%	100.0%	66.7%	100.0%
L1 University Text	Precision	87.5%	80.0%	62.5%	83.8%	100.0%	73.3%	90.9%
	Recall	72.4%	66.7%	100.0%	88.6%	100.0%	68.8%	83.3%
Short Story	Precision	86.8%	75.0%	61.5%	72.4%	100.0%	80.0%	60.7%
	Recall	64.7%	75.0%	80.0%	75.0%	85.7%	100.0%	89.5%

Legend	40.0%	55.0%	70.0%	85.0%	100.0%

*Table A-3: Precision and recall scores for each thematic progression pattern from the **test** dataset. This breakdown illustrates the degree to which parsing errors permeated thematic progression classification in terms of correct classification (precision) and extraction of thematic progression patterns present in the texts (recall). Higher precision and recall scores (highlighted in green) indicate a higher overall parsing accuracy.*

Bibliography

Texts Used for Training and Testing

1804 dollar. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/1804_dollar (5 May, 2022).

2007–2008 Nazko earthquakes. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/2007%E2%80%932008_Nazko_earthquakes (5 May, 2022).

2013 Atlantic hurricane season. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/2013_Atlantic_hurricane_season (5 May, 2022).

Acacia pycnantha. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Acacia_pycnantha (5 May, 2022).

Acamptonectes. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Acamptonectes (5 May, 2022).

Achelousaurus. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Achelousaurus (5 May, 2022).

Acra. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Acra_%28fortress%29 (5 May, 2022).

Actuary. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Actuary (5 May, 2022).

Adele. 2021. Lyrics to "Dew on the Vine". https://www.azlyrics.com/lyrics/adele/easyonme.html.

AE Systems. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/BAE_Systems (5 May, 2022).

African crake. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/African_crake (5 May, 2022).

After the Deluge. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/After_the_Deluge_(painting) (5 May, 2022).

Agaricus deserticola. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Agaricus_deserticola (5 May, 2022).

Akodon spegazzinii. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Akodon_spegazzinii (5 May, 2022).

Aleeta curvicosta. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Aleeta_curvicosta (5 May, 2022).

American paddlefish. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/American_paddlefish (5 May, 2022).

Anas, Brittany. 2022. 31 Little Ways to Save for a Down Payment This Month. https://www.apartmenttherapy.com/31-ways-save-for-down-payment-37071641 (5 May, 2022).

Anonymous. 2022. *Afghanistan:* University Essay. Friedrich-Alexander University.

Anonymous. 2009. *A CVX theory of Dutch Syllable Structure.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Augustine Anatomized.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Autoethnography.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Case Study Analysis of Emergency Contraception Can Reveal Characteristics of Society as a Whole.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Economics of the Illicit-Drug Market.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Educational Autobiography.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Fetal Endocrine System.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Ideology and the Transition to Democracy.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Individual Colonies: Dialect Acquisition in Immigrants.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Jim Colbert Final Paper.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Leave for Family Responsibility in the United States.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).

Anonymous. 2009. *Lesson reflection (class on cloud types).* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Manufacturing System Design for a Car Assembly Process.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *ME 513 Mini-Project Report.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Mix Proportioning and Fresh Concrete Properties.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Moral Approval and Disapproval.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *My Life Is Not a Movie Starring Michelle Pfeiffer or Hilary Swank.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Psychology Underground of Power: Memo #1.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Rubberland Liquid Rubber Failure Analysis.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Sleep Patterns in Infancy as a Predictor of Insomnia in College Students.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Springer and Breines.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Teleportation with Trapped Ions.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *The American Landscape: A Personal Reflection From a Bike Seat.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *The Evolution of Terrestriality: A Look at the Fcators that Drove Tetrapods to Move onto Land.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *The Social Impact of Roe v. Wade.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *The Vicar of Wakefield as a Failed Morality Story.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *The World Bank and IMF: Broken but Worth Fixing.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Week 6: New Social History (Historicizing Race, Religion, and Culture).* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *Why did the U.S. develop a health care system that relies mostly in the private provision of health insurance?.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2009. *World Leaders Pretend.* In: The Michigan Corpus of Upper-level Student Papers (MICUSP).
Anonymous. 2022. *3D Printing.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Additive Manufacturing.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Armenian Genocide.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Bitcoin.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Chess.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Competitive Labor Market.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Computed Tomography.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Cover Letter.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Covid-19 Microvascular Complications.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Ctenophora.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Enduring Freedom.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *German Code of Criminal Procedure.* University Essay. Friedrich-Alexander University.

Anonymous. 2022. *Graph Description.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *MENA-Region.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Mental Health.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Modern World.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *NATO Membership.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Ne bis in idem.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *One Road Initiative.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Personality Types.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Psychopaths.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *School Closures.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Shareholders.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Ted Bundy.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *US Unemployment Rate.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Usable Farmland.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Video Games.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Women in India.* University Essay. Friedrich-Alexander University.
Anonymous. 2022. *Xenotransplanation.* University Essay. Friedrich-Alexander University.
Archaea. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Archaea (5 May, 2022).
Atlantic Star. 1987. Lyrics to "Always". https://www.lyrics.com/lyric/35069279/Atlantic+Starr.
Atom and His Package. 1997. Lyrics to "Atom and His Package". https://genius.com/Atom-and-his-package-atom-and-his-package-lyrics.
Barr, Leticia. 2022. Comcast Laptop Donation to Easterseals Keeps Communities Connected. https://techsavvymama.com/2022/01/comcast-laptop-donation-to-easterseals.html (5 May, 2022).
Bear's Den. 2016. Lyrics to "Dew on the Vine". https://www.lyrics.com/lyric/33135177/Dew+on+the+Vine.
Belghast. 2022. Sad Rockboy is Sad. https://aggronaut.com/2022/04/20/sad-rockboy-is-sad/ (5 May, 2022).
Berlin to Kitchener name change. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Berlin_to_Kitchener_name_change (5 May, 2022).
Bird, Andrew. 2007. Lyrics to "Armchairs". https://genius.com/Andrew-bird-armchairs-lyrics.
Bon Jovi. 1988. Lyrics to "Bad Medicine". https://www.azlyrics.com/lyrics/bonjovi/badmedicine.html.
Bronwyn Bancroft. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Bronwyn_Bancroft (5 May, 2022).
Brown, Brene. 2021. Brené Brown's Atlas of the Heart: Defensiveness and Flooding. https://www.gottman.com/blog/brene-browns-atlas-of-the-heart-defensiveness-and-flooding/ (5 May, 2022).
Christie. 1970. Lyrics to "Yellow River". https://www.lyrics.com/lyric/15334225/Christie/Yellow+River.
Coelho, Camila. 2022. Why use aluminum-free deodorant? https://camilacoelho.com/2022/04/27/why-use-aluminum-free-deodorant/ (5 May, 2022).
Collins, Phil. 1988. Lyrics to "A Groovy Kind of Love". https://www.lyrics.com/lyric/33118176/Phil+Collins/A+Groovy+Kind+of+Love.
Death Cab for Cutie. 2022. Lyrics to "Roman Candles". https://www.musixmatch.com/lyrics/Death-Cab-for-Cutie/Roman-Candles.
Demira. 2023. Lyrics to "Fountain". https://genius.com/Demira-fountain-lyrics (5 May, 2022).
Dunlop, Katie. 2018. 3 Days in Aspen. https://lovesweatfitness.com/3-days-in-aspen/ (5 May, 2022).
Fantasy Book. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Fantasy_Book (5 May, 2022).
Gayle. 2021. Lyrics to "abcdefu". https://genius.com/Gayle-abcdefu-lyrics.
Goddard, Joanna. 2019. My Grandmother's Home in Cornwall. https://cupofjo.com/2019/11/05/cornwall-house-tour/ (5 May, 2022).
Hand Luggage Only. 2022. 14 Very Best Things To Do in Zakynthos, Greece. https://handluggageonly.co.uk/2022/02/08/14-very-best-things-to-do-in-zakynthos-greece/ (5 May, 2022).

Harlow, Jack. 2022. Lyrics to "First Class". https://genius.com/Jack-harlow-first-class-lyrics.
Harold Melvin and the Blue Notes. 1975. Lyrics to "Don't Leave Me This Way". https://www.lyrics.com/lyric/3738254/The+Communards/Don%27t+Leave+Me+This+Way.
Heick, Terry. 2023. A Self-Directed Learning Model For 21st Century Learners. https://www.teachthought.com/critical-thinking/self-directed-learning-model- (5 May, 2022).
Helen in Wanderlust. 2022. The Best Things To Do in Sierra Leone: A 2-Week Itinerary. https://www.heleninwonderlust.co.uk/best-things-to-do-in-sierra-leone/ (5 May, 2022).
Ian, Janis. 1975. Lyrics to "At Seventeen". https://www.lyrics.com/lyric/29627667/Janis+Ian/At+Seventeen.
Johnson, Nicole. 2016. Toddler Won't Stay in Bed? 6 Must-Know Tips for Toddler Sleep Regression. https://www.rookiemoms.com/5-tips-when-your-toddler-wont-stay-in-bed-and-sleep/ (5 May, 2022).
Kamb, Steve. 2022. The Star Wars Workout: Begin Your Jedi Training! https://www.nerdfitness.com/blog/the-star-wars-workout-jedi-training-101/ (5 May, 2022).
Kay, Maisy. 2022. Lyrics to "Karma is a Bitch Like You". https://genius.com/Maisy-kay-karma-is-a-bitch-like-you-lyrics.
Kellogg, Kathryn. 2020. 10 Easy Swaps for Plastic Free July. https://www.goingzerowaste.com/blog/easy-swaps-for-plastic-free-july/ (5 May, 2022).
Kirschner, Diana. 2022. Sudden Break Up of a Long Term Relationship: 5 Best Hacks. https://lovein90days.com/sudden-break-up-long-term-relationship-5-best-hacks/ (5 May, 2022).
Lefebvre, Eliot. 2022. Vague Patch Notes: The art of picking winners and losers in MMO design. https://massivelyop.com/2022/11/03/vague-patch-notes-the-art-of-picking-winners-and-losers-in-mmo-design/ (5 May, 2022).
Lizzo. 2022. Lyrics to "About Damn Time". https://www.songtexte.com/songtext/lizzo/about-damn-time-g6b49b2ae.html.
Maroon 5. 2012. Lyrics to "Payphone". https://www.lyrics.com/lyric/37302074/Wiz+Khalifa/Payphone.
Mary Anning. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Mary_Anning (5 May, 2022).
Natalie. 2022. Cricut EasyPress 3 Project – Personalized Kid Aprons. https://www.athomewithnatalie.com/cricut-easypress3-project-personalized-kid-aprons/ (5 May, 2022).
Omar, Danielle. 2022. How to Sneak More Veggies Into Your Favorite Family Recipes. https://blog.myfitnesspal.com/how-to-sneak-more-veggies-into-your-favorite-family-recipes/ (5 May, 2022).
Pallett, Owen. 2006. Lyrics to "Song Song Song". https://genius.com/Owen-pallett-song-song-song-lyrics.
Pendle Witches. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Pendle_witches (5 May, 2022).
Plan B. 2010. Lyrics to "She Said". https://www.lyrics.com/lyric/36111229/She+Said.
Radiohead. 2000. Lyrics to "Idioteque". https://www.lyrics.com/lyric/18639720/Radiohead/Idioteque+%5BLive%5D.
Ronie, Baden. 2022. Ravenous Devils Review – The Demon Tailor of Fleet Street. https://wolfsgamingblog.com/2022/04/29/ravenous-devils-review-the-demon-tailor-of-fleet-street/ (5 May, 2022).
Rose. 2021. Lyrics to "On the Ground". https://genius.com/Rose-on-the-ground-lyrics.
Ross, Diana. 1980. Lyrics to "Upside Down". https://www.azlyrics.com/lyrics/dianaross/upsidedown.html.
Saunders, Paul T. 2022. Lyrics to "Appointment in Samarra". http://www.lyricshall.com/lyrics/Paul+Thomas+Saunders/Appointment+In+Samarra/.
Seattle Center Monorail. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Seattle_Center_Monorail (5 May, 2022).
September 1964 South Vietnamese coup attempt. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/September_1964_South_Vietnamese_coup_attempt (5 May, 2022).

Sharma, Romit. 2022. 7 Doable Ways To Manage Remote Teams. https://www.techcrackblog.com/2022/05/ways-to-manage-remote-teams.html (5 May, 2022).
SLAPP Suits. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/SLAPP_Suits (5 May, 2022).
Snavely, Andrew. 2022. The Amazon Outfit: Wear This to a Last-minute Wedding You Put Off Shopping For. https://www.primermagazine.com/2022/spend/last-minute-wedding-outfit-men (5 May, 2022).
Solar System. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Solar_System (5 May, 2022).
Stanley, Liz. 2022. A Peaceful Primary Bathroom Renovation. https://sayyes.com/2022/03/a-peaceful-primary-bathroom-renovation (5 May, 2022).
Styles, Harry. 2022. Lyrics to "As It Was". https://genius.com/Harry-styles-as-it-was-lyrics.
Taylor, Kathryne. 2020. Southwestern Corn Chowder. https://cookieandkate.com/vegetarian-corn-chowder-recipe/comment-page-1/ (5 May, 2022).
Taylor, Will. 2022. Modern and Rustic Guest Bathroom. https://www.brightbazaarblog.com/2022/01/modern-and-rustic-guest-bathroom.html (5 May, 2022).
Terry, Beth. 2019. Spring Gardening with Orta Plastic-Free Self-Watering Seed Pots. https://myplasticfreelife.com/2019/03/spring-gardening-with-orta-plastic-free-self-watering-seed-pots/ (5 May, 2022).
The Bat. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/The_Bat_(play) (5 May, 2022).
The Boat Race 2012. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/The_Boat_Race_2012 (5 May, 2022).
The Book of Kells. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Book_of_Kells (5 May, 2022).
The Box Tops. 2013. Lyrics to "The Letter". https://www.lyrics.com/lyric/29627731/The+Box+Tops.
The Everly Brothers. 2006. Lyrics to "Cathy's Clown". https://www.lyrics.com/lyric/8683855/The+Everly+Brothers.
The Kid LAROI & Justin Bieber. 2020. Lyrics to "Stay". https://www.azlyrics.com/lyrics/kidlaroi/stay.html.
The Smitten Kitchen Digest. 2022. Simplest mushroom pasta. https://smittenkitchen.com/2022/04/simplest-mushroom-pasta/ (5 May, 2022).
Tracey, A. J. 2021. Lyrics to "Little More Love". https://genius.com/Aj-tracey-little-more-love-lyrics.
Travels of Adam. 2022. How to Be the Best Airbnb Host…from a Guest's Perspective. https://travelsofadam.com/how-to-airbnb-host/ (5 May, 2022).
Tyga. 2021. Lyrics to "Lift Me Up". https://www.lyrics.com/lyric-lf/6710544/Tyga.
Velarmino, Trisha. 2021. The unfortunate truths of starting an online business: what we don't talk about. https://www.psimonmyway.com/cons-of-starting-an-online-business/ (5 May, 2022).
Virginia Eliza Clemm Poe. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Virginia_Eliza_Clemm_Poe (5 May, 2022).
Vora, Nisha. 2022. Vegan Lilac Lemon Cake. https://rainbowplantlife.com/vegan-lilac-lemon-cake/#wprm-recipe-container-5659 (5 May, 2022).
Waddington, Elizabeth. 2022. Edible Landscaping Tips for Beginners. https://www.treehugger.com/edible-landscaping-tips-beginners-5248277 (5 May, 2022).
Wainwright, Rufus. 2004. Lyrics to "The Art Teacher". https://www.lyrics.com/lyric/27758551/Rufus+Wainwright.
Walters, Jennipher & Kristen Seymour. 2008-2023. Fit Bottomed Girls: About. https://fitbottomedgirls.com/about-us/ (5 May, 2022).
William Barley. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/William_Barley (5 May, 2022).
Yurko, Eric. 2022. #881 – Factory Funner [BGT Edition]. https://whatsericplaying.com/2022/04/04/factory-funner-bgt-edition/ (5 May, 2022).

Primary Literature

Adam, Martin. 2013. *Presentation sentences*: *(syntax, semantics and FSP)* (Spisy Pedagogické fakulty Masarykovy univerzity 162). Brno: Masaryk University.

Agichtein, Eugene, Walt Askew & Yandong Liu. 2008. Combining Lexical, Syntactic, and Semantic Evidence for Textual Entailment Classification. *Theory and Applications of Categories* 31. 1–6.

Ahmed, Mahmoud A., Mohamed Eltayeb Abdalla & Ayman Hamed Elneil Hamdan. 2015. The Impact of Thematization & Contextualiztion as Discoursal Features on the Quality of EFL M.A. Students' Written Performance. *SUST Journal of Humanities* 16(4). 266–278.

Ammann, Hermann. 1928. *Die menschliche Rede. II.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Arnold, Jennifer, Elsi E. Kaiser, Jason M. Kahn & Lucy K. Kim. 2013. Information structure: linguistic, cognitive, and processing approaches. *Wiley interdisciplinary reviews. Cognitive science* 4(4). 403–413.

Aumüller, Matthias. 2014. Text Types. *Handbook of Narratology*. 854–867.

Austin, John L., James O. Urmson & Marina Sbisa. ca. 2009 = 1975. *How to do things with words*: *The William James lectures delivered at Harvard University in 1955,* 2nd edn. Cambridge, Mass.: Harvard Univ. Press.

Badger, Richard. 2003. Legal and general: towards a genre analysis of newspaper law reports. *English for Specific Purposes* 22(3). 249–263.

Bakhtin, Michail M. 2010. *Speech genres and other late essays* (University of Texas Press Slavic series 8), 12th edn. Austin, Tex.: Univ. of Texas Press.

Berry, Margaret. 1995. Thematic options and success in writing. In Mohsen Ghadessy (ed.), *Thematic Development in English Texts*, 55–84.

Biber, Douglas. 1995. *Variation across speech and writing,* 1st edn. Cambridge: Cambridge University Press.

Biber, Douglas, Ulla Connor & Thomas A. Upton. 2007. *Discourse on the move*: *Using corpus analysis to describe discourse structure* (Studies in corpus linguistics v. 28). Amsterdam, Philadelphia: John Benjamins Pub. Co.

Bohnet, Bernd, Alicia Burga & Leo Wanner. 2013. Towards the Annotation of Penn TreeBank with Information Structure. In *Proceedings of the sixth international joint conference on natural language processing.* 1250–1256.

Boyatzis, Richard E. 2010. *Transforming qualitative information*: *Thematic analysis and code development.* Thousand Oaks, Calif.: SAGE Publications Ltd.

Breivik, Leiv E. 1981. On the Interpretation of Existential There. *Language* 57(1). 1–25.

Burlaga, Christine M. 2004. *A contrastive approach to the thematic analysis of text and genre: An examination of lead news articles in Le Monde, Al-Ittihad, and The New York Times.* California State University Master Thesis.

Chai, Haixia & Michael Strube. 2022. Incorporating Centering Theory into Neural Coreference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2996–3002.

Chamonikolasová, Jana & Martin Adam. 2005. *The Presentation Scale in the Theory of Functional Sentence Perspective.* Prague: Department of English and American Studies, Faculty of Arts, Charles Univ.

Chen, Xiaobin, Theodora Alexopoulou & Ianthi Tsimpli. 2021. Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behavior Research Methods* 53(2). 803–817.

Cloran, Carmel. 1995. Defining and relating text segments: Subject and Theme in Discourse. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4. 361–403.

Collins, Peter C. 2015. *Cleft and Pseudo-Cleft Constructions in English* (Routledge Library Edition: The English Language Volume 6). Hoboken: Taylor and Francis.

Connor, Ulla & Thomas Upton. 2003. Linguistic Dimensions of Direct Mail Letters. In Pepi Leistyna (ed.), *Corpus analysis*: *Language structure and language use* (*Language and computers* 46), 71–86. Amsterdam: Rodopi.
Crossley, Scott A., Laura K. Varner, Rod D. Roscoe & Danielle S. McNamara. 2013. Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013.* 269–278.
Daneš, Frantisek. 1974. *Papers on functional sentence perspective* (Janua Linguarum. Series Minor Ser v.147). Berlin/Boston: De Gruyter.
Davidse, Kristin. 1987. M.A.K. Halliday's functional grammar and the Prague school. In Vilém Fried & René Dirven (eds.), *Functionalism in Linguistics*, 39–79. Amsterdam: John Benjamins Publishing Company.
Davies, Florence. 1997. Marked theme as a heuristic for analysing text-type, text and genre. *Applied languages: theory and practice in ESP*. 45–79.
Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.
Derczynski, Leon. 2016. Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 261–266.
Derewianka, Beverly. 2011. *A New Grammar Companion for Teachers.* Newtown, N.S.W.: Primary English Teaching Association.
Derewianka, Beverly. 2012. Knowledge about language in the Australian curriculum: English. *Faculty of Social Sciences - Papers (Archive)*. 127–146.
Domínguez, Monica, Juan Soler & Leo Wanner. 2020. ThemePro: A Toolkit for the Analysis of Thematic Progression. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1000–1007.
Dou, Chen & Shuo Zhao. 2018. An Analysis of Themes and Thematic Progression Patterns in Ivanka Trump's Speech. *Studies in Literature and Language* 16(3). 62–67.
Downing, Angela. 2001. Thematic progression as a functional resource in analysing texts. *CLAC (Circulo de Lingüistica Aplicada a la Communicacion)* 5(2). 23-42.
Downing, Angela & Philip Locke. 2006. *English Grammar,* 2nd edn. Milton: Taylor & Francis Group.
Dubois, Betty L. 1987. A reformulation of thematic progression typology. *Text - Interdisciplinary Journal for the Study of Discourse* 7(2). 89–116.
Ebrahimi, Seyed F. 2016. Theme Types and Patterns in Research Article Abstracts: A Cross Disciplinary Study. *International Journal of English Language and Translation Studies*. 104–115.
Eggins, Suzanne. 2004. *An introduction to systemic functional linguistics.* London: Continuum.
Enkvist, Nils E. 1973. "Theme dynamics" and style: An experiment. *Studia Anglica Posnaniensia* 5. 127–135.
Enkvist, Nils E. 1987. A Note Towards the Definition of Text Strategy. *STUF - Language Typology and Universals* 40(1-6). 19–27.
Ferret, Olivier & Brigitte Grau. 1998. A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts. In *13th European Conference on Artificial Intelligence*. 155-159.
Figueiredo, Débora. 2010. Context, register and genre: Implications for language education. *Revista signos* 43. 119–141.
Firbas, Jan. 1964a. From comparative word-order studies: Thoughts on V. Mathesius' conception of the word-order system in English compared with that in Czech. *Brno Studies in English* 4. 111–126.
Firbas, Jan. 1964b. On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague* 1. 267–280.
Firbas, Jan. 1986. On the dynamics of written communication in the light of the theory of functional sentence perspective. *Studying writing: Linguistic approaches*. 40–71.
Firbas, Jan. 1987. On the delimitation of the theme in functional sentence perspective. In Vilém Fried & René Dirven (eds.), *Functionalism in Linguistics*, 137–156. Amsterdam: John Benjamins Publishing Company.

Firbas, Jan. 1992. *Functional sentence perspective in written and spoken communication* (Studies in English language). Cambridge: Cambridge University Press.
Firbas, Jan. 1996. Mobility of clause constituents and functional sentence perspective. In Eva Hajičová, Barbara Hall Partee & Petr Sgall (eds.), *Discourse and meaning*: *Papers in honor of Eva Hajičová*, 221–233. Amsterdam, Philadelphia: J. Benjamins.
Forey, Gail. 2002. *Aspects of theme and their role in workplace texts:* University of Glasgow Doctoral Thesis.
Francis, Gill. 1989. Thematic selection and distribution in written discourse. *Word* 40(1-2). 201–221.
Frey, Werner. 2003. Syntactic conditions on adjunct classes. In Cathrine Fabricius-Hansen, Ewald Lang & Claudia Maienborn (eds.), *Modifying adjuncts* (*Interface explorations* 4), 163–210. Berlin/Boston: De Gruyter Mouton.
Fries, Peter H. 1981. On the status of theme in English: Arguments from discourse. *Forum Linguisticum* 6. 1–38.
Fries, Peter H. 1992. The structuring of information in written English text. *Language Sciences* 14(4). 461–488.
Fries, Peter H. 1995. Themes, methods of development, and texts. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4. 317–360.
Fu, Qiankun, Linfeng Song, Wenyu Du & Yue Zhang. 2021. End-to-End AMR Coreference Resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4204–4214.
Gerot, Linda & Peter Wignell. 1994. *Making sense of functional grammar.* Cammeray, NSW: Antipodean Educational Enterprises.
Gholamy, Afshin, Vladik Kreinovich & Olga Kosheleva. 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*. 1–6.
Gómez, María A. 1994. The relevance of Theme in the textual organization of BBC news reports. *Word* 45(3). 293–305.
Gomez-Licon, Adriana. 2022. Floods trap many in Florida as Ian heads to South Carolina. *The Detroit News*.
Gosden, Hugh. 1992. Discourse functions of marked theme in scientific research articles. *English for Specific Purposes* 11(3). 207–224.
Gosden, Hugh. 1994. *A genre-based investigation of Theme: Product and process in scientific research articles written by NNS novice researchers:* University of Liverpool Doctoral Thesis.
Green, Christopher F., Elsie R. Christopher & Jaquelin Lam Kam Mei. 2000. The Incidence and Effects on Coherence of Marked Themes in Interlanguage Texts: A Corpus-Based Enquiry. *English for Specific Purposes* 19(2). 99–113.
Guijarro, Arsenio J. M. & Jesús Ángel Ávila Zamorano. 2009. Thematic progression of children's stories as related to different stages of cognitive development. *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies* 29(6). 755–774.
Gunawan, Wawan & Fatayatul Aziza. 2017. Theme and Thematic Progression of Undergraduate Thesis: Investigating Meaning Making in Academic Writing. *Indonesian Journal of Applied Linguistics* 7(2). 413–424.
Gundel, Jeanette K. 1988. Universals of topic-comment structure. In *Studies in Syntactic Typology*, 209–240: John Benjamins.
Hajicová, Eva & Jiří Mírovský. 2018. Discourse coherence through the lens of an annotated text corpus: A case study. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 1637–1642.
Hajičová, Eva & Jarka Vrbová. 1981. On the salience of the elements of the stock of shared knowledge. *Folia Linguistica,* Available at: https://doi.org/10.1515/flin.1981.15.3-4.291.
Halliday, Michael A. K. 1967. Notes on transitivity and theme in English. *Journal of Linguistics* 3(2). 199–244.

Halliday, Michael A. K. 1970. Language structure and language function. *New horizons in linguistics* 1. 140–165.
Halliday, Michael A. K. 1974. The place of 'functional sentence perspective' in the system of linguistic description. In *Papers on functional sentence perspective*, 43–53: De Gruyter.
Halliday, Michael A. K. 1985. *An introduction to functional grammar,* 1st edn. London: Edward Arnold.
Halliday, Michael A. K. & Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar,* 3rd edn. London: Routledge.
Halliday, Michael A. K. & Christian M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar 4th edition,* 4th edn. Hoboken: Taylor and Francis.
Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English* (English language series no. 9). London, New York: Taylor and Francis.
Hasselgård, Hilde. 2010. *Adjunct adverbials in English* (Studies in English language). Cambridge, UK: Cambridge University Press.
Hasselgård, Hilde. 2020. In the Case of Theme: Topic Identifiers in English and Norwegian Academic Texts. *Contrastive Pragmatics* 1(1). 108–135.
Hawes, Thomas. 2001. *Thematisation in the editorials of The Sun and The Times:* University of Liverpool Unpublished Doctoral Thesis.
Hawes, Thomas. 2010a. Breaks in Thematic Progression. *Philologia* 8(1). 31–45.
Hawes, Thomas. 2010b. Thematic progression and rhetoric in Sun and Times editorials: 1991-2008. *Rice Working Papers in Linguistics* 2. 39–51.
Hawes, Thomas. 2015. Thematic progression in the writing of students and professionals. *Ampersand* 2. 93–100.
Hawes, Thomas & Sarah Thomas. 1997a. Problems of Thematisation in Student Writing. *RELC Journal* 28(2). 35–55.
Hawes, Thomas & Sarah Thomas. 1997b. Rhetorical uses of theme in newspaper editorials. *World Englishes* 15(2). 159–170.
Hawes, Thomas & Sarah Thomas. 2012. Theme choice in EAP and media language. *Journal of English for Academic Purposes* 11(3). 175–183.
Herriman, Jennifer. 2011. Themes and theme progression in Swedish advanced learners' writing in English. *Nordic Journal of English Studies* 10(1). 1–28.
Hinkel, Eli. 2001. Matters of Cohesion in L2 Academic Texts. *Applied Language Learning* 12(2). 111–132.
Ho, Debbie G. E. 2017. Thematic Options and Success in ESL Writing: An Analysis of Promotional Texts. In Peter Mickan & Elise Lopez (eds.), *Text-Based Research and Teaching*, 281–304. London: Palgrave Macmillan UK.
Hoey, Michael. 2005. *Lexical priming*: *A new theory of words and languages.* London: Routledge.
Honnibal, Matthew, Ines Montani, Sofie van Landeghem & Adriane Boyd. 2020a. *spaCy*: *Industrial-strength Natural Language Processing in Python*. Available at: https://spacy.io.
Honnibal, Matthew, Ines Montani, Sofie van Landeghem & Adriane Boyd. 2020b. Spacy Doc.noun_chunks: spaCy: Industrial-strength Natural Language Processing in Python. Available at: https://spacy.io/api/doc#noun_chunks.
Honnibal, Matthew, Ines Montani, Sofie van Landeghem & Adriane Boyd. 2020c. Spacy Word vectors and semantic similarity: spaCy: Industrial-strength Natural Language Processing in Python. Available at: https://spacy.io/usage/linguistic-features#vectors-similarity.
Honnibal, Matthew, Ines Montani, Sofie van Landeghem & Adriane Boyd. 2020d. Facts and Figures: spaCy: Industrial-strength Natural Language Processing in Python. Available at: https://spacy.io/usage/facts-figures.
Hotho, Andreas, Steffen Staab & Gerd Stumme. 2003. Ontologies improve text document clustering. In *Third IEEE International Conference on Data Mining*. 541–544.
Huang, Jing & Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*. 917–920.

Huddleston, Rodney, Geoffrey K. Pullum & Brett Reynolds. 2021. *A Student's Introduction to English Grammar.* Cambridge: Cambridge University Press.
Hudson, Richard P. 2022. *Coreferee:* Explosion AI. Available at: https://github.com/msg-systems/coreferee
Israel, Ross, Joel Tetreault & Martin Chodorow. 2012. Correcting Comma Errors in Learner Essays, and Restoring Commas in Newswire Text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 284–294.
Jalilifar, Alireza. 2009. Thematic Development in English and Translated Academic Texts. *Journal of Universal Language* 10(1). 81–111.
Jalilifar, Alireza. 2010. Thematization in EFL Students' Composition Writing and Its Relation to Academic Experience. *RELC Journal: A Journal of Language Teaching and Research* 41(1). 31–45.
Jalilifar, Alireza & Ebtesam Abbasi Montazeri. 2017. Thematicity in Applied Linguistics Textbooks: A Comparative Study of Foreword, Introduction and Preface. *Iranian Journal of Language Teaching Research* 5(2). 15–36.
Jingxia, Liu & Li Liu. 2013. An Empirical Study on the Application of Theme Theory in the Field of Writing Pedagogy. *English Language Teaching* 6(5). 117–128.
Jonge, Casper C. de. 2007. From Demetrius To Dik. Ancient And Modern Views On Greek And Latin Word Order. In *The Language of Literature*, 211–232: BRILL.
Joty, Shafiq, Giuseppe Carenini & Raymond T. Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics* 41(3). 385–435.
Kamath, Aishwarya, Jonas Pfeiffer, Edoardo Maria Ponti, GoranGlavaš & Ivan Vulić. 2019. Specializing Distributional Vectors of All Words for Lexical Entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 72–83.
Kanoksilapatham, Budsaba. 2007. Rhetorical moves in biochemistry research articles. In Douglas Biber, Ulla Connor & Thomas A. Upton (eds.), *Discourse on the move*: *Using corpus analysis to describe discourse structure* (*Studies in corpus linguistics* 28), 73–119. Amsterdam, Philadelphia: John Benjamins Pub. Co.
Kappagoda, Astika. 2009. *The Use of Systemic-Functional Linguistics in Automated Text Mining.* Australia.
Kazemian, Bahram & Somayyeh Hashemi. 2014. Critical Discourse Analysis of Barack Obama's 2012 Speeches: Views from Systemic Functional Linguistics and Rhetoric. *Theory and Practice in Language Studies* 4(6). 1178–1187.
Khurana, Diksha, Aditya Koli, Kiran Khatter & Sukhdev Singh. 2022. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications* 82(3). 3713–3744.
Kopple, William J. V. 1991. Themes, Thematic Progressions, and Some Implications for Understanding Discourse. *Written Communication* 8(3). 311–347.
Kress, Gunter. 2014. Genre as Social Process. In Bill Cope & Mary Kalantzis (eds.), *The powers of literacy*: *A genre approach to teaching writing* (Routledge Library Editions: Education v. 113), 22–37. Hoboken: Taylor and Francis.
Lancia, Franco. 2012. T-lab Pathways to Thematic Analysis. Available at: http://www.mytlab.com/tpathways.pdf.
Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar.* Stanford, CA: Stanford University Press.
Lavid, Julia L. 2000. Linguistic and computational approaches to information in discourse: theme, focus, given, and other dangerous things. In *Revista Canaria de Estudios Ingleses* 40. 355–369.
Lee, David Y. 2001. Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. In Bernhard Kettemann & Georg Marko (eds.), *Teaching and Learning by Doing Corpus Analysis*, 245–292: Amsterdam: Rodopi.

Lemke, Jay L. 1983. Thematic Analysis: Systems, Structures and Strategies. *Recherches Semiotiques Semiotic Inquiry*. 159–187.
Lemke, Jay L. 1994. Genre as a Strategic Resource. Available at: https://eric.ed.gov/?id=ED377515.
Leong, Ping A. 2000. Identifying the Theme of Existential Clauses: A Suggested Approach. *Folia Linguistica* 34(3-4). 777–780.
Leong, Ping A. 2005. Talking themes: the thematic structure of talk. *Discourse Studies* 7(6). 701–732.
Leong, Ping A. 2007. Developing the Message: Thematic Progression and Student Writing. *The Journal of AsiaTEFL* 4(3). 93–127.
Leong, Ping A. 2015. Topical themes and thematic progression: the "picture" of research articles. *Text & Talk* 35(3). 289–315.
Leong, Ping A. 2019. Visualizing texts: a tool for generating thematic-progression diagrams. *Functional Linguistics* 6(1). 1–13.
Leong, Ping A., Audrey L.L. Toh & Soo Fun Chin. 2018. Examining Structure in Scientific Research Articles: A Study of Thematic Progression and Thematic Density. *Written Communication* 35(3). 286–314.
Leopold, Edda & Jörg Kindermann. 2002. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning* 46(1/3). 423–444.
Li, Qing-Feng. 2009. Thematic selection and progression in EFL writing. *US-China Foreign Language* 7(7). 25-28.
Loftipour-Saedi, Kazem & Forouzan Rezai-Tajani. 1996. Exploration in thematization strategies and their discoursal values in English. *Text - Interdisciplinary Journal for the Study of Discourse* 16(2). 225–250.
Lorés, Rosa. 2004. On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes* 23(3). 280–302.
Lu, Xiaofei. 2002. Discourse and ideology: The Taiwan issue in the Chinese and American media. *Research and practice in professional discourse*. 589–608.
Ma, Jianjun & Jiaqi Zhu. 2023. A Comparative Study on Promotion of Modal Adjuncts in Research Article Introductions. *Theory and Practice in Language Studies* 13(2). 463–472.
Martin, Jim R. 1984. Language, register and genre. In Frances Christie (ed.), *Children Writing: Reader*, 21–30. Geelong, AU: Deakin University Press.
Martin, Jim R. 1985. Process and text: two aspects of human semiosis. *Systemic Perspectives on Discourse* 1. 248–274.
Martin, Jim R. 1992. *English Text.* Amsterdam: John Benjamins Publishing Company.
Martin, Jim R. 1993. A Contextual Theory of Language. *The powers of literacy: A genre approach to teaching writing*. 116–136.
Martin, Jim R. 1995. More than what the message is about: English Theme. In Mohsen Ghadessy (ed.), *Thematic development in English texts* (Open linguistics series), 223–259. London: Pinter.
Martínez, Iliana A. 2003. Aspects of theme in the method and discussion sections of biology journal articles in English. *Journal of English for Academic Purposes* 2(2). 103–123.
Mathesius, Vilém. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis.* Prague: Academia.
Mathesius, Vilém. 1983. Functional Linguistics. In Josef Vachek & Libuše Dušková (eds.), *Praguiana* (12), 121–142. Amsterdam: John Benjamins Publishing Company.
Mathesius, Vilém. 1985. New Currents and Tendencies in Linguistic Research. In Endre Bojtar (ed.), *Slavic Structuralism* (*Linguistic and Literary Studies in Eastern Europe* 11), 45. Amsterdam: John Benjamins Publishing Company.
Matthiessen, Christian M. I. M. 1995. Theme as an enabling resource in ideational 'knowledge' construction. In Mohsen Ghadessy (ed.), *Thematic development in English texts* (Open linguistics series). London: Pinter.
Mauranen, Anna. 1993. *Cultural differences in academic rhetoric*: *A textlinguistic study* (Nordeuropäische Beiträge aus den Human- und Gesellschaftswissenschaften 4). Frankfurt am Main: Peter Lang.

McCabe, Anne M. 1999. *Theme and thematic patterns in Spanish and English history texts:* Aston University Doctoral Thesis.
McCabe, Anne M. & Isabel Alonso Belmonte. 1998. Theme-Rheme patterns in L2 writing. *Didáctica* 10. 13–31.
Michigan Corpus of Upper-Level Student Papers. 2009. Ann Arbor, MI: The Regents of the University of Michigan.
Moens, Marie-Francine. 2007. Using patterns of thematic progression for building a table of contents of a text. *Natural Language Engineering* 14(2). 145–172.
Moore, Nick A. J. 2006. Aligning Theme and Information Structure to Improve the Readability of Technical Writing. *Journal of Technical Writing and Communication* 36(1). 43–55.
Morton, Donald. 2011. Problem and Solution. Available at: https://www.ereadingworksheets.com/text-structure/patterns-of-organization/problem-and-solution/.
Muttaqin, Muhammad. 2017. *Thematic analysis of spoken texts in the English dialogue: a study at the VIII grade of SMP N 16 Semarang in academic year 2016/2017.* Walisongo State Islamic University Bachelor Thesis.
Naderi, Sahar & Farhang Koohestanian. 2014. Thematic Structures in Conference Papers by Persian EFL Scholars. *Procedia - Social and Behavioral Sciences* 118. 351–356.
Nivre, Joakim. 2022. Universal Dependencies. Available at: https://universaldependencies.org/#language-.
North, Sarah. 2005. Disciplinary Variation in the Use of Theme in Undergraduate Essays. *Applied Linguistics* 26(3). 431–452.
Park, Kwanghyun & Xiaofei Lu. 2015. Automatic analysis of thematic structure in written English. *International Journal of Corpus Linguistics* 20(1). 81–101.
Patten, Amanda. 2012. *The English It-Cleft: A Constructional Account and a Diachronic Investigation* (Topics in English Linguistics [TiEL]). Berlin, Boston: De Gruyter Mouton.
Plotly Technologies Inc. 2015. *Collaborative data science.* Montréal, QC. Available at: https://plotly.com/.
Plum, Guenter A. 1988. *Text and Contextual Conditioning in Spoken English: A Genre-based Approach:* University of Sydney Doctoral Thesis.
Popping, Roel. 2000. *Computer-Assisted Text Analysis*: SAGE Publications, Ltd, https://doi.org/10.4135/9781849208741
Prince, Ellen F. 1981. Towards a taxonomy of given-new information. In Cole, P. (ed.), *Radical Pragmatics*. 223–255. New York, NY: Academic Press.
Pritchard, Elena. 2012. Brand Awareness: Family Branding vs. Individual Branding. Available at: http://performancemarketingtoday.blogspot.com/2012/03/brand-awareness-family-branding-vs.html
Puşcaşu, Georgiana, Patricio Martínez Barco & Estela Saquete Boró. 2006. On the Identification of Temporal Clauses. In *MICAI 2006: Advances in Artificial Intelligence: 5th Mexican International Conference on Artificial Intelligence*. 911–921.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language,* 24th edn. London, New York: Longman.
Řehůřek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.
Řehůřek, Radim & Petr Sojka. 2011. *Gensim-python framework for vector space modelling.* NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).
Repka, Richard. 2021. The Prague School of Linguistics and Halliday's Systemic and Functional Grammar. *Philologia* 31(1). 165–177.
Roller, Stephen, Douwe Kiela & Maximilian Nickel. 2018. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In Iryna Gurevych & Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 358–363.

Rørvik, Sylvi. 2012. Thematic progression in learner language. In Sebastian Hoffmann, Paul Rayson & Geoffrey N. Leech (eds.), *English corpus linguistics*: *Looking back, moving forward: papers from the 30th International Conference on English Language Research on Computerized Corpora (ICAME 30*). 165–177.
Rosa, Rusdi N. 2013. Thematic progression as a means to keep cohesion in exposition text. In *Proceedings of ISELT FBS Universitas Negeri Padang* 1. 220–228.
Rose, David. 2001. Some variations in Theme across languages. *Functions of Language* 8(1). 109–145.
Rudolph, Udo & Friedrich Försterling. 1997. The psychological causality implicit in verbs: A review. *Psychological Bulletin* 121(2). 192–218.
Saikh, Tanik, Sudip Kumar Naskar, Chandan Giri & Sivaji Bandyopadhyay. 2015. Textual Entailment Using Different Similarity Metrics. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015*. 491–501.
Sanner, Michel F. 1999. Python: a programming language for software integration and development. *Journal of molecular graphics & modelling* 17(1). 57–61.
Sarkar, Dipanjan. 2019. *Text Analytics with Python*: *A Practitioner's Guide to Natural Language Processing,* 2nd edn. Berkeley, CA: Apress; Imprint: Apress.
Saussure, Ferdinand de. 1988. Chapter 6. Langue/Parole. In Robert M. Strozier (ed.), *Saussure, Derrida, and the Metaphysics of Subjectivity* (Approaches to Semiotics [AS] 80). Berlin: De Gruyter.
Scarpa, Federica. 2020. *Research and Professional Practice in Specialised Translation* (Palgrave Studies in Translating and Interpreting), 1st edn. London: Palgrave Macmillan UK; Palgrave Macmillan.
Scharkow, Michael. 2013. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity* 47(2). 761–773.
Schwarz, Lara, Sabine Bartsch, Richard Eckart & Elke Teich. 2008. Exploring automatic theme identification: a rule-based approach. In Annely Rothkegel, John Laffling, Angelika Storrer, Alexander Geyken, Alexander Siebert & Kay-Michael Würzner (eds.), *Text Resources and Lexical Knowledge* (*Text, Translation, Computational Processing*), 15–26. Berlin, New York: Mouton de Gruyter.
Sgall, Petr, Eva Benesova & Eva Hajičová. 1973. *Topic, focus and generative semantics* (Forschungen Linguistik und Kommunikationswissenschaft 1). Kronberg: Scriptor.
Shi, Jian. 2013. The Exploration of the Topical Progression Patterns in English Discourse Analysis. *Theory and Practice in Language Studies* 3(9). 1639–1644.
Shum, Buckingham S., Ágnes Sándor, Rosalie Goldsmith, Randall Bass & Mindy McWilliams. 2017. Towards Reflective Writing Analytics: Rationale, Methodology and Preliminary Results. *Journal of Learning Analytics* 4(1). 58–84.
Simsek, Duygu, Simon Buckingham Shum, Agnes Sandor, Anna de Liddo & Rebecca Ferguson. 2013. XIP Dashboard: visual analytics from automated rhetorical parsing of scientific metadiscourse. In *1st International Workshop on Discourse-Centric Learning Analytics*. 265–266.
Sinclair, John & Ronald Carter. 2014. *Trust the Text*: *Language, Corpus and Discourse,* 1st edn. Florence, Ann Arbor, Michigan: Taylor and Francis; ProQuest.
Singh, Sameer, Dustin Hillard & Chris Leggetter. 2010. Minimally-Supervised Extraction of Entities from Text Advertisements. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 73–81.
Smolka, Vladislav. 2017. What Comes First, What Comes Next: Information Packaging in Written and Spoken Language. A*cta Universitatis Carolinae Philologica* 1. 51–61.
Søgaard, Anders, Miryam de Lhoneux & Isabelle Augenstein. 2018. Nightmare at test time: How punctuation prevents parsers from generalizing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 25–29.
Steinberger, Ralf & Paul Bennett. 1994. Automatic Recognition of Theme, Focus and Contrastive Stress. In *Proceedings of the Conference on Focus and Natural Language Processing*. 22–34.

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie & Ellen Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Vol. 2)*. 656–664.

Stylianou, Nikolaos & Ioannis Vlahavas. 2021. A neural Entity Coreference Resolution review. *Expert Systems with Applications* 168. 1–20.

Svoboda, Aleš. 1981. *Diatheme: A Study in Thematic Elements, Their Contextual Ties, Thematic Progressions and Scene Progressions Based on a Text from Aelfric*. Brne: Univerzita J.E. Purkyne.

Swales, John M. 1990. *Genre analysis: English in academic and research settings.* Cambridge, UK: Cambridge Univ. Press.

Taglicht, Josef. 1984. *Message and emphasis: On focus and scope in English.* London: Longman.

Thelin, Ryan. 2021. Python Version History: How Python has changed over the years. Available at: https://www.educative.io/blog/python-versions-history.

Thompson, Geoff. 2004. *Introducing functional grammar.* London: Arnold.

van Atteveldt, Wouter, Mariken A. C. G. van der Velden & Mark Boukes. 2021. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures* 15(2). 121–140.

van Dijk, Teun A. 1977. Sentence topic and discourse topic. *Papers in slavic philology*(1). 49–61.

Vian Jr., Orlando & Re de Lima-Lòpes. 2005. A perspectiva teleológica de Martin para a análise dos gêneros textuais. *Generos: teorias, metodos, debates*. 29–45.

Vilnis, Luke & Andrew McCallum. 2015. Word Representations via Gaussian Embedding. In *Proceedings of the 2015 International Conference on Learning Representations.*

Vulić, Ivan & Nikola Mrkšić. 2018. Specialising Word Vectors for Lexical Entailment. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* 1134–1145.

Weil, Henri. 1978. *The order of words in the ancient languages compared with that of the modern languages* (Amsterdam Classics in Linguistics, 1800–1925 v.14). Amsterdam, Philadelphia: John Benjamins B.V.

Wiedemann, Gregor. 2016. Computer-Assisted Text Analysis in the Social Sciences. In *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. 17–54.

Williams, Ian A. 2009. Discourse style and theme–rheme progression in biomedical research article discussions. *Languages in Contrast* 9(2). 225–266.

Winter, Eugene. 1982. *Towards a Contextual Grammar of English.* London: George Allen and Unwin.

Witte, Stephen P. 1983. Topical Structure and Revision: An Exploratory Study. *College Composition and Communication* 34(3). 313–341.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

Xi, Xuefeng, Victor S. Sheng, Shuhui Yang, Baochuan Fu & Zhiming Cui. 2020. Predicting Simplified Thematic Progression Pattern for Discourse Analysis. *Computers, Materials & Continua* 62(3). 163–181.

Yuned, Reski O. 2016. Coherence analysis of the 2015 international conference article abstracts in applied linguistics. In *Proceedings of ISELT FBS Universitas Negeri Padang* 4(2). 199–209.

Zahra, Galis M., Emi Emilia & Iyen Nurlaelawati. 2021. An Analysis of Cohesion and Coherence of Descriptive Texts Written by Junior High School Students. *Thirteenth Conference on Applied Linguistics (CONAPLIN 2020)*. 195–202.