LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
OPEN PUBLISHING IN THE HUMANITIES
UB

AXEL WISIOREK

Quantitative Methoden einer kognitiven Texttypologie

Automatische Genre-Klassifizierung als Rekonstruktion kognitiver Weltmodelle

GEORG OLMS VERLAG

Quantitative Methoden einer kognitiven Texttypologie

Automatische Genre-Klassifizierung als Rekonstruktion kognitiver Weltmodelle

Inauguraldissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität München

vorgelegt von
Axel Wisiorek
aus Freising
2023

Referent: PD Dr. Peter-Arnold Mumm
Korreferent: Prof. Dr. Thomas Krefeld
Datum der mündlichen Prüfung: 19.02.2021

Open Publishing in the Humanities

In der Reihe Open Publishing in the Humanities (OPH) wird die Veröffentlichung von hervorragenden geistes- und sozialwissenschaftlichen Dissertationen gefördert. Die LMU unterstützt damit Open Access als *best practice* in der Publikationskultur von Monografien in den Geistes- und Sozialwissenschaften und engagiert sich zugleich in der Nachwuchsförderung. Herausgeber von OPH sind Prof. Dr. Hubertus Kohle und Prof. Dr. Thomas Krefeld.

Die Universitätsbibliothek der LMU stellt dafür ihre Infrastruktur des hybriden Publizierens bereit und ermöglicht dadurch jungen, forschungsstarken WissenschaftlerInnen, ihre Werke gedruckt und gleichzeitig auch Open Access zu veröffentlichen.

https://oph.ub.uni-muenchen.de

Axel Wisiorek

Quantitative Methoden einer kognitiven Texttypologie
Automatische Genre-Klassifizierung als Rekonstruktion kognitiver Weltmodelle

Quantitative Methoden einer kognitiven Texttypologie

Automatische Genre-Klassifizierung als Rekonstruktion kognitiver Weltmodelle

von
Axel Wisiorek

GEORG OLMS VERLAG

Universitätsbibliothek
Ludwig-Maximilians-Universität München

Eine Publikation in Zusammenarbeit zwischen dem **Georg Olms Verlag** und der **Universitätsbibliothek der LMU München**

Gefördert von der Ludwig-Maximilians-Universität München

Erstveröffentlichung 2023
Zugleich Dissertation der LMU München 2021

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet abrufbar über http://dnb.d-nb.de

Open-Access-Version dieser Publikation verfügbar unter:
http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-108214-3
https://doi.org/10.5282/oph.19

ISBN 978-3-487-16116-7

Inhaltsverzeichnis

Vorwort

Vorliegende Arbeit, die auf Anregung von Prof. Dr. Wolfgang Schulze entstand, erweitert seine von ihm im Rahmen einer Text-Weltmodell-Theorie entwickelte Methodik für eine gebrauchsbasierte Genre-Typologie um automatische Klassifizierungs- und Mustererkennungsverfahren der Computerlinguistik. In der korpusbasierten Modellierung kognitiver Texttypen durch Methoden des maschinellen Lernens, die im Zentrum der Arbeit steht, verbinden sich meine sprachwissenschaftlichen ideal mit meinen computerlinguistischen Forschungsinteressen, sodass ich den Vorschlag meines Doktorvaters, dieses Thema zum Gegenstand meiner Dissertation zu machen, begeistert aufgenommen habe.

Als Datengrundlage für die Anwendung der in meiner Arbeit entwickelten Verfahren einer automatischen Genre-Klassifizierung konnte ich auf die innerhalb des DFG/FWF-geförderten Projekts „Ob-Ugric Database“ (OUDB) informationsstrukturell annotierten Korpora zurückgreifen, an deren Aufbau ich als Projektmitarbeiter mit beteiligt war. Dr. Zsófia Schön und Dr. Gwen Eva Janda, die diese primär mündlichen, z. T. in eigener Feldforschung gewonnenen Sprachdaten der sibirischen Sprachgemeinschaften der Khanten und Mansen im Rahmen des von Prof. Dr. Elena Skribnik geleiteten OUDB-Projekts aufbereitet haben, danke ich für die gute Zusammenarbeit und hoffe gleichzeitig, mit der vorliegenden Arbeit auch etwas zur Erforschung der obugrischen Informationsstruktur beitragen zu können.

Als Wolfgang Schulze einige Monate vor der geplanten Einreichung meiner Dissertation im Frühjahr 2020 unerwartet verstarb, verlor ich mit ihm nicht nur meinen Doktorvater, sondern auch einen Mentor, der für mich als Mensch und als Sprachwissenschaftler stets ein großes Vorbild gewesen ist. Wolfgang hat immer an mich geglaubt, und ich bin sehr dankbar, ihn gekannt und zum Doktorvater gehabt zu haben. Umso mehr hoffe ich, mit dieser Arbeit sein Werk in seinem Sinne ein kleines Stück weiterzuführen. Seinem Andenken widme ich diese Arbeit.

Mein ganz besonderer Dank gilt PD Dr. Peter-Arnold Mumm, der nach Wolfgangs Tod, ohne zu zögern, die Betreuung meiner Dissertation übernommen hat und mir durch seine Unterstützung in dieser schwierigen Zeit so den erfolgreichen Abschluss meiner Promotion ermöglichte. Ganz herzlich möchte ich mich auch bei Prof. Dr. Thomas Krefeld für die Übernahme des Zweitgutachtens, bei PD Dr. Ilona Schulze für das Drittgutachten sowie bei Prof. Dr. Hinrich Schütze für die Beteiligung am Promotionsverfahren bedanken, ebenso für all die Begleitung und den Beistand, den ich im Verlauf meiner Promotion erfahren habe.

Den Herausgebern Prof. Dr. Hubertus Kohle und Prof. Dr. Thomas Krefeld danke ich für die Aufnahme in die von der Ludwig-Maximilians-Universität München geförderte Reihe „Open Publishing in the Humanities“ der Universitätsbibliothek München.

Ohne die tatkräftige Unterstützung und Betreuung der Publikation durch die Mitarbeiter*innen der Publikationsdienste Open Access des Referats Elektronisches Publizieren der Universitätsbibliothek München Dr. Claudie Paye, Andrea Dorner, Annerose Wahl und Dan-Mihai Pitea sowie das Lektorat von Karin Hennig hätte ich die Veröffentlichung dieses Buchs in dieser Form nicht realisieren können. Dafür danke ich ihnen sehr herzlich.

München, im November 2023 Axel Wisiorek

English Summary

Quantitative Methods of a Cognitive Text Typology: Automatic Genre Classification as a Reconstruction of Cognitive World Models

This study proceeds from the basic cognitive-linguistic assumption that the structuring of texts as sequences of speech acts follows genre-specific rules and patterns forming part of the cognitive knowledge system of a speaker within a language community, namely in the form of schematic structural models (text world models, Schulze 2018; 2019; 2020). Learned through cognitive processes of conventionalization and classification via the typification of recurrent structural patterns in language use, these abstract, usage-based cognitive models (Johnson-Laird 1983; Langacker 2000) regulate the production and reception of texts of a given genre (Miller 1984). This involves adapting the structure of the individual cognitive text models constructed during text processing (situation models, van Dijk & Kintsch 1983) to the specific type of communicative situation in which the texts arise.

When applied as cognitive structural rules in text production, the genre-specific schemata laid down in a text world model establish a trace (Langacker 2000; Schwarz 2000) within the linguistic structure of a text. This study will explore statistical classification methods for identifying such genre-typical text structure patterns (Heinemann & Viehweger 1991; Fix 2008) based on recurring, quantitative features in texts of a given genre. The text grammars (van Dijk 1972) resulting from the extraction of prototypical structural feature values from corpora as language usage data can then ultimately facilitate the reconstruction of the schematic rules of the corresponding text world models by means of interpreting their relational, referential and information-structural characteristics as parameters of a cognitive text linguistics (Schulze 2018; 2019; 2020).

For this intended corpus-based exploration of cognitive genre models, this study relies on quantitative text structure parameters such as information density, elaboration measures or frequent event patterns that can be assumed to be relevant to the construction of cognitive text models, with their selection being primarily based on the taxonomy of parameters of text world models (TWM parameters) as developed by Schulze (2018; 2019; 2020). To operationalize these, feature construction methods of representing texts via both multivariate text-structural feature sets and sequence-based text structure patterns are introduced, including their extraction from annotated text corpora using data mining techniques. Appropriate clustering and classification methods for the exploratory study of such text structure representations are proposed, including Random Forest classifiers and Dynamic Time Warping-based clustering, which eventually allow for a quantitative cognitive text typology based on TWM parameters.

In a concluding pre-test study for such a cognitively grounded genre classification, the established methods for representing and classifying texts based on TWM parameters as linguistically encoded, schematic patterns for constructing genre-specific cognitive text models are tested on an information-structurally annotated, historically and dialectally stratified corpus of Ob-Ugrian folk tales as well as texts from other genres. Since these texts of oral tradition are close to the original language practice (Schulze 2019), they are well suited as test cases for the intended usage-based identification of genres and subgenres as text structure pattern types which can be interpreted as culture-specific orders of discourse (Foucault 1981; 1991; Schulze 2020). As the study shows, several such types of structural organization can indeed be recognized in the Ob-Ugric corpus via the proposed TWM operationalization.

1 Einleitung

1.1 Theoretische Grundlagen und Terminologie

1.1.1 Modellbasierte Informationsverarbeitung

Gemäß dem Grundparadigma der kognitiven Psychologie verarbeitet der Mensch die Welt als in der Wahrnehmung gegebene Sachverhalte über sein informationsverarbeitendes System der Kognition durch Konstruktion von **kognitiven Modellen** dieser Sachverhalte.[1] Ein kognitives Modell (Johnson-Laird 1983: *mental model*; van Dijk & Kintsch 1983: *situation model*; vgl. Zwaan & Radvansky 1998) ist eine interne, gedankliche Repräsentation eines wahrgenommenen (oder sprachlich oder anderweitig vermittelten) Ausschnitts von Wirklichkeit im informationsverarbeitenden System des Menschen. Im Fall sprachlicher Vermittlung ist das konstruierte Modell „the cognitive representation of the events, actions, persons, and in general the situation, a text is about" (van Dijk & Kintsch 1983: 11f.). Die in linear aufeinander abfolgenden Einzelwahrnehmungen als zeit-räumlich strukturiert[2] gegebenen Sachverhalte werden durch strukturerhaltende, aber komplexitätsreduzierende Abbildungen in mentale Repräsentationen transformiert.[3] Die so konstruierten kognitiven Modelle spiegeln also zentrale Relationen und Zusammenhänge zwischen den in einer Situation beteiligten Objekten modellhaft (vgl. Häcker 2020: 863; Kopp & Caspar 2020: 1550f.; Kopp 2020: 1551; Gigerenzer 2020: 939f.; Schulze 2004a: 546f.), d. h. mit einer reduzierten, aber zu den repräsentierten Sachverhalten analogen Struktur:

1 Kognition wird hier im Sinne der kognitiven Psychologie aufgefasst als Gesamtheit der informationsverarbeitenden mentalen Prozesse eines Individuums, die der Manipulation und Speicherung von Informationen als mentalen Repräsentationen im Gedächtnis dienen, insbesondere der von Wissenseinheiten im Langzeitgedächtnis, welche wiederum Grundlage für verhaltenssteuernde Prozesse sind (vgl. Gigerenzer 2020: 939f.).
2 Vgl. Raum und Zeit als Formen der menschlichen Wahrnehmung bei Kant: „Vermittelst des äußeren Sinnes […] stellen wir uns Gegenstände als außer uns, und diese insgesamt im Raume vor. […] [A]llein es ist doch eine bestimmte Form, unter der die Anschauung ihres innern Zustandes allein möglich ist, so, daß alles, was zu den innern Bestimmungen gehört, in Verhältnissen der Zeit vorgestellt wird." (Kant 1998: 97, B37)
3 Die Komplexitätsreduktion der zu verarbeitenden Informationen ist aufgrund der eingeschränkten Kapazität des Arbeitsgedächtnisses notwendig; dieses wird nach dem klassischen Komponentenmodell als der Teil des Kurzzeitgedächtnisses aufgefasst, der die sequentiell in der Wahrnehmung gegebenen Informationen mit den permanent im Langzeitgedächtnis abgespeicherten Informationen abgleicht (Bredenkamp 2020: 660f.; Seitz 2020: 186). Nach neueren, prozessorientierten Ansätzen (Embedded-Processes-Modell, s. Dreisbach 2020: 480) ist das Arbeitsgedächtnis der Teil des Gedächtnisses, dessen Informationen sich im Zustand der Aktivierung befinden, d. h. dem Bewusstsein zugänglich und damit verarbeitbar sind (Dreisbach 2020: 480); die Verarbeitung dieser aktivierten Informationen des Arbeitsgedächtnisses wird über die Aufmerksamkeitssteuerung, d. h. über Prozesse zur Selektion von Informationen, reguliert (Krummenacher 2020: 221f.), wobei die Informationskapazität des Bereichs des Fokus der Aufmerksamkeit auf einen Umfang von nur wenigen (potentiell komplexen) Informationseinheiten (sog. Chunks, s. Dreisbach 2020: 480) beschränkt ist (s. van Dijk 2018: 29; Bredenkamp 2020: 660f.; Gigerenzer 2020: 939f.).

> A model represents a state of affairs and accordingly its structure is not arbitrary like that of a propositional representation, but plays a direct representational or analogical role. Its structure mirrors the relevant aspects of the corresponding state of affairs in the world. (Johnson-Laird 1980: 98)

Gemäß der Schematheorie der kognitiven Psychologie (Bartlett 1932; Minsky 1974) basieren die Kriterien für die in der Modellkonstruktion durch die Kognition eines Individuums stattfindende Informationsreduktion auf dessen bisheriger Erfahrung ähnlicher Situationen (s. Kopp 2020: 1551; vgl. Brewer & Nakamura 1984; Mandl & Friedrich & Hron 1988: 124ff.). Der die Komplexitätsreduktion gewährleistende kognitive Filterungsprozess wählt dementsprechend die für eine Situation typischen Strukturzusammenhänge für die Konstruktion des kognitiven Modells dieser Situation aus.[4] Was die für eine Situation typischen Zusammenhänge sind, wird durch im kognitiven System verankerte Mustererkennungsprozesse bestimmt, die Regelmäßigkeiten in der Struktur der über die Wahrnehmung gegebenen Informationen feststellen; solche Muster – also wiederholt auftretende Strukturen – werden in der kognitiven Psychologie als **Schemata** bezeichnet:

> Schemata sind übergeordnete kogn. Strukturen von Gegenständen, Situationen und Inhalten, die das Verstehen gewährleisten, indem neu wahrgenommene Informationen einem adäquaten Schema zugeordnet werden. Zugleich werden dadurch die neuen Informationen für das kogn. System zugänglich, abrufbar und erweiterbar gemacht (Informationsverarbeitung). Damit einher geht ein Reduktionsprozess: Umfangreiche Informationen werden an den adäquaten Stellen hinzugefügt bzw. zu einer übergeordneten Struktur zus.gefasst. (Kopp 2020: 1551)

Im Sinne der Theorie mentaler Modelle kann man Schemata – also „Wissensstrukturen, in denen aufgrund von Erfahrungen typische Zusammenhänge eines Realitätsbereichs repräsentiert sind“ (Mandl & Friedrich & Hron 1988: 124; vgl. Brewer & Nakamura 1984: 7) – als mentale Modelle höheren Abstraktionsgrads auffassen (*higher level conceptualizations*, Rumelhart & Ortony 1977: 109f.; vgl. Brewer & Nakamura 1984: 27f.; Mandl & Friedrich & Hron 1988: 151). Solche **abstrakten Strukturmodelle**, die die Konstruktion (d. h. die schematische Strukturierung) von Situationsmodellen steuern, werden durch Typisierung der Erfahrung ähnlicher Situationen erlernt, indem wiederholt (regelmäßig) auftretende Objekte und Abläu-

4 Diese Strategie der Verarbeitung neuer Situationen, basierend auf typischen Eigenschaften und regelhaften Zusammenhängen bekannter Situationen, die in kognitiven Kategorisierungs- und Klassifikationsprozessen bestimmt werden, ist notwendig aufgrund der Kapazitätsbeschränkung des menschlichen Arbeitsgedächtnisses – gleichzeitig ist diese reduktionistische Strategie aber auch ökonomisch, da Situationen so schnell und (aufgrund der Erwartbarkeit regelhafter Abläufe in der Welt) zumeist erfolgreich verarbeitet werden können (vgl. van Dijk 2018: 30; DeLamater & Myers & Collett 2018: 210).

fe abgespeichert werden – repräsentiert etwa über Schemavariablen, die bei der auf diesem Strukturschema basierenden Konstruktion eines kognitiven Modells durch den jeweiligen Kontext gefüllt werden (Kopp 2020: 1551); so etwa bei Skripts als über Rollen-Leerstellen realisierte schematische Repräsentationen bestimmter typischer Abfolgen von Ereignissen und Handlungen (vgl. Schwarz-Friesel & Consten 2014: 71).

Kognitive Schemata sind also Typisierungen von Situationen; sie repräsentieren als Struktur-Prototypen (vgl. DeLamater & Myers & Collett 2018: 208) stereotype strukturelle Eigenschaften einer Situationsklasse, indem sie von der Vielzahl der in den Einzelwahrnehmungen einer Situation gegebenen Informationen zu den beteiligten Objekten abstrahieren (s. Schwarz 2000: 34; vgl. Kopp & Caspar 2020: 1550f.).

Die Konstruktion von kognitiven Situationsmodellen wird durch dem jeweiligen Situationstyp entsprechende, im Langzeitgedächtnis als Wissenseinheiten abgespeicherte Strukturmodelle höheren Abstraktionsgrads gesteuert, die die zentralen Struktureigenschaften bisher erfahrener, ähnlich strukturierter Situationen abbilden, so eine Kategorisierung von neuen Informationen im Rahmen bekannter Modelle ermöglichen und auf diese Weise „das Wahrgenommene […] strukturieren" (Kopp 2020: 1551). Kognitive Situationsmodelle haben demnach eine grundsätzlich schematische Struktur – d. h. eine auf zentrale (d. h. aufgrund der Erfahrung ähnlicher Situationen relevante) Kategorien reduzierte Struktur; sie sind „temporäre analoge Repräsentationen, die in Abhängigkeit von den bisherigen Erfahrungen des Individuums auf der Grundlage kogn. Schemata konstruiert werden und deshalb jew. typ. Sachverhalte repräsentieren (Prototyp)" (Häcker 2020: 863).

Als Grundlage der schemabasierten Verarbeitung von Informationen der Welt kann man die je nach Situationstyp aktivierten, abstrakten Strukturmodelle, die Wissen über typische Strukturzusammenhänge von Bereichen der Welt repräsentieren, auch als **Weltmodelle** bezeichnen (Schulze 2020: 605; s. auch Schulze 2018: 172; 2019: 7):[5] Ein Weltmodell ist demnach ein abstraktes Modell der Struktur eines Situationstyps, das induktiv durch Generalisierung von wiederholten Erfahrungen ähnlicher Situationen durch Kategorisierungs- und Lernprozesse in der Kognition aufgebaut (Schemainduktion) bzw. angepasst wird (Mustervergleich) und im Gedächtnis als Wissenseinheit abgespeichert ist (s. Kopp & Caspar 2020: 1550f.). Neue Situationen können so mit Hilfe eines Weltmodells aufgrund der bereits gemachten Erfahrung ähnlicher Situationen klassifiziert und entsprechend verarbeitet werden. Bei der Verarbeitung neuer Wahrnehmungen von Sachverhalten wird dann ein dieser Situation adäquates Weltmodell aktiviert (d. h. aus dem Gedächtnis abgerufen), das die Verarbeitung (Strukturierung) der in der Wahrnehmung gegebenen Informationen

5 Zur Textlage einer Weltmodell-bezogenen Theorie der Textverarbeitung bei Schulze s. auch Abschnitt 1.2, Fußnote 15.

durch sukzessive Konstruktion des entsprechenden Situationsmodells leitet; dazu gehört beispielsweise die Erwartung von bestimmten Strukturpositionen (Leerstellen), die im jeweiligen Kontext mit konkreten Werten gefüllt werden (Schwarz-Friesel & Consten 2014: 71). Die Verarbeitung durch Weltmodelle ermöglicht auch **Inferenz** bzw. Elaboration, d. h. in der Wahrnehmung fehlende Informationen werden durch das im situativ aktivierten Schema abgespeicherte Wissen um die auftretenden Objekte und Ereignisse ergänzt (s. DeLamater & Myers & Collett 2018: 212). Das durch schematische Strukturmodelle konstruierte kognitive Modell einer Situation hat also einerseits eine auf zentrale kognitive Kategorien reduzierte (schematische) Struktur, ist aber andererseits durch die Ergänzung von Informationen aus mit der Situation verbundendem Schemawissen komplexer (dies gilt insbesondere für textuell kodierte Modelle; s. van Dijk 2016: 67). Entsprechend steuern diese erlernten Weltmodelle auch das Verhalten und Handeln des Individuums in der Welt (vgl. Piaget 1948):

> Planning of behavior requires some knowledge about the consequences of actions in a given environment. A *world model* captures such knowledge. There is clear evidence that nervous systems use such internal models to perform predictive motor control, imagery, inference, and planning in a way that involves a simulation of actions and their perceptual implications [...]. (Toussaint 2003: 929, Hervorhebung im Original)

1.1.2 Genrespezifische Sprachverarbeitung

Nach einer Grundannahme der kognitiven Textlinguistik wird auch Text – hier definiert als die in einer Kommunikationssituation geäußerte Sequenz von Sprachhandlungen – analog zu anderem in der Wahrnehmung gegebenem Umwelt-Input von der menschlichen Kognition über ein der jeweiligen Situation adäquates schematisches Strukturmodell verarbeitet, das hier mit Schulze als **Text-Weltmodell** (**TWM**) bezeichnet wird (Schulze 2018: 172ff.; 2019: 7ff.; 2020: 604ff.).

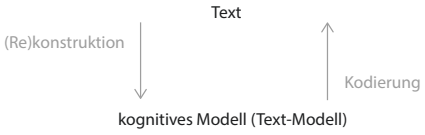


Abbildung 1.1: Kodierung und (Re)konstruktion eines kognitiven Text-Modells

Während in den Text-Welt-Theorien von Werth (1999), Schwarz (2000; 2001; 2011) oder Gavins (2007) unter einem *Text-Welt-Modell* das konkrete, Einzeltext-bezogene kognitive Modell verstanden wird, das bei Produktion bzw. Rezeption eines Textes durch den Sprecher[6] bzw. Hörer konstruiert wird – d. h.

6 Das in dieser Arbeit zumeist gewählte generische Maskulinum bezieht sich zugleich auf die männliche, die weibliche und andere Geschlechteridentitäten. Zur besseren Lesbarkeit wird auf die Verwendung männlicher und weiblicher Sprachformen verzichtet. Alle Geschlechteridentitäten werden ausdrücklich mitgemeint, soweit die Aussagen dies erfordern.

also das spezifische „mentale Sachverhaltsmodell“ (Schwarz-Friesel & Consten 2014: 73) der in einem Text kodierten ‚Welt‘ (im Folgenden als **Text-Modell** bezeichnet)[7] – bezieht sich Schulze mit dem Begriff des Text-*Weltmodells* (Schulze 2019: 8f.; vgl. Schulze 2020: 629) auf ein abstraktes kognitives Modell, das als schematisches Strukturmodell den Aufbau von Text-Modellen eines ähnlichen Situationstyps (eines Genres, s. u.) in ihrer Produktion bzw. Verarbeitung steuert (s. auch 2.1.3). Schulze bezieht sich mit diesem Begriff also auf ein TWM höheren Abstraktionsgrades, nämlich das kognitive Modell eines Typs von ‚Textwelten‘; vgl. dazu Schulze 2019: 8 (Hervorhebung im Original): „Während *Textwelt-Modell* paraphrasiert werden kann als ‚Modell der Bedeutung eines Textes‘ […], steht *Text-Weltmodell* für ‚textuell repräsentierte Weltmodelle‘.“

Anders als physikalische oder biologische Sachverhalte haben Sprachhandlungen als soziale Prozesse eine in einer Sprechergemeinschaft durch Übereinkunft etablierte, also konventionelle Verweisfunktion: In der zwischenmenschlichen Interaktion erzeugen Menschen mit Sprachhandlungen für andere wahrnehmbare Sachverhalte in der Welt (vgl. Schulze 2019: 5), die auf mentale Repräsentationen von realen oder fiktiven Sachverhalten verweisen (darin liegt ihre Bedeutung als sprachliche Zeichen). Diese Repräsentationen sind also kognitiv mit dem jeweiligen Ausdruck verknüpft und werden bei dessen Wahrnehmung aktiviert; diese Verknüpfung wird in der Sprachpraxis gelernt und als Sprachwissen abgespeichert.

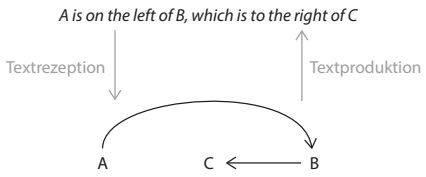


Abbildung 1.2: Textbasiertes kognitives Modell (reproduziert nach Johnson-Laird 1983: 164)

Dem klassischen Sender-Empfänger-Modell entsprechend dient also die in einer Kommunikationssituation geäußerte Sequenz von Sprachhandlungen der Übermittlung eines kognitiven Situationsmodells als mentaler Repräsentation einer Situation von Sprecher zu Hörer (vgl. DeLamater & Myers & Collett 2018: 272). Auch ein solches Text-Modell als ein über einen Text kodiertes kognitives Modell besitzt – wie oben festgestellt – **schematische** Struktur: Es beinhaltet also kognitive Kategorien wie raum-zeitliche Verortung, Referenten (d. h. sprachlich kodierte Objekte) sowie Ereignisse (sprach-

7 Vgl. die TWM-Definition von Schwarz-Friesel & Consten 2014: 58: „Das TWM [*Textweltmodell*] stellt eine rein geistige Zwischenebene im Arbeitsgedächtnis (während oder kurz nach der Rezeption) bzw. im Langzeitgedächtnis (auch nach der Rezeption) dar, die durch die Informationseinheiten des Textes aufgebaut wird und Referenten als mentale Einheiten mit ihren Relationen und Aktivitäten sowie ihrer raumzeitlichen Verankerung speichert.“

lich kodierte Relationszusammenhänge zwischen den Referenten).[8] Abbildung 1.2 zeigt die graphische Darstellung eines einfachen textbasierten kognitiven Modells nach einem Beispiel von Johnson-Laird (1983: 164).

Das schematische Weltmodell, das als Grundlage für die Konstruktion eines solchen kognitiven Text-Modells dient, wird induktiv durch Typisierung der Erfahrung ähnlicher Kommunikationssituationen erlernt (s. 1.1.1; vgl. auch Miller 1984; Schütz & Luckmann 2003) bzw. z. T. auch durch direkte Vermittlung über didaktischen oder expositorischen Diskurs angeeignet (van Dijk 2018: 34):

> Zusammenfassend kann also ein Text-Weltmodell verstanden werden als die konzeptuelle Basis eines Inventars von rhetorischen Handlungsmustern, die unter anderem durch Erfahrung und soziales Lernen erworben wird und verankert ist in den kollektiven Wissensraum einer Gemeinschaft, mithin ein Teil der gegebenen soziokulturellen und praxeologischen Normen und Konventionen ist. (Schulze 2019: 9)

In einer Kommunikationssituation wird dann das zur aktuellen Situation passende Text-Weltmodell zur Steuerung des Strukturaufbaus sowohl in der Text-Produktion als auch in der Text-Rezeption aktiviert; diese erlernten Texttypen sind also **Typisierungen** von rekurrenten Kommunikationssituationen der spezifischen Lebenswelt (Schütz & Luckmann 2003) einer Sprechergemeinschaft.[9]

Eine solche Kategorisierung von Sprachhandlungssequenzen nach ihrem Kommunikationskontext als gebrauchsbasierte Texttypen kann man mit diskurstypologischen und sozialpsychologischen Ansätzen (etwa Swales 1990; Miller 1984; s. auch Bawarshi & Reiff 2010) unter dem Begriff des **Genres** fassen: „The notion of 'genre' designates a conventional way of performing communicative activities using language" (Stukker & Spooren & Steen 2016: 1). Miller definiert Genre als „conventional category of discourse based in large-scale typification of rhetorical action; as action, it acquires meaning from situation and from the social context in which that situation arose" (Miller 1984: 163). Ein Text-Weltmodell ist als abstraktes kognitives Modell

8 Vgl. van Dijk 2018: 30: „Since we observe, participate or talk about events many thousands of times in our lives, mental models have a standard schematic structure of a limited number of categories that allow very fast processing, probably developed during human evolution, such as Setting (Place, Time), Participants (and their Identities, Roles and Relationships), Event or Action (and its Intention or Purpose). Such a schema allows us to 'analyse' and understand a situation or event in fractions of seconds and then take appropriate action, as is also the case in conversation."

9 S. Schütz & Luckmann 2003: 318ff. zu „Typik und Sprache": „Die Sprache ist ein System typisierender Erfahrungsschemata, das auf Idealisierungen und Anonymisierungen der unmittelbaren subjektiven Erfahrung beruht. […] Diese von der Subjektivität abgelösten Erfahrungstypisierungen sind sozial objektiviert, wodurch sie zu einem Bestandteil des dem Subjekt vorgegebenen gesellschaftlichen Apriori werden. Für den normalen erwachsenen Menschen in der natürlichen Einstellung ist deshalb Typisierung aufs engste mit der Sprache verschränkt." (Schütz & Luckmann 2003: 318)

eines bestimmten Typs schematischer Sprachhandlungsstruktur (Textstruktur) also die mentale Repräsentation des dadurch kodierten Genres:

> Dieses [*kognitive Schema*] repräsentiert die Organisationsstruktur eines Textes. Schemata sind ganzheitliche Repräsentationen von best. Ereigniszusammenhängen. Das bekannteste Schema ist die sog. *story grammar* (Rumelhart, 1975). Sie repräsentiert die Organisationsstruktur eines Erzähltextes. (Engelkamp 2020: 1781, Hervorhebung im Original; vgl. Rumelhart 1975)

Im Sinne der Konstruktionsgrammatik (Goldberg 1995) kann man auch einem Text-Weltmodell (als abstraktem kognitiven Konstrukt) Zeichencharakter zusprechen (Schulze 2019: 4f.). So wie ein Text im Sinne der Textlinguistik (vgl. Figge 2000) als komplexes sprachliches Zeichen ein einzelnes Situationsmodell kodiert, kodiert die Struktur eines Textes dasjenige Text-Weltmodell, das als abstraktes schematisches Strukturmodell des Texttyps (Genres) die Strukturierung von Texten dieses Typs steuert. Bestimmte rekurrente textstrukturelle Merkmale (Textstrukturmuster, vgl. Heinemann & Viehweger 1991: 170; Fix 2008: 67) repräsentieren also genrespezifische, im Langzeitgedächtnis abgespeicherte TWM-Modell-Parameter (vgl. Schulze 2019: 4).

In diesem Sinne kann man ein Genre als komplexes konstruktionelles Zeichen auffassen (s. Schulze 2019: 6, 13), das ein Text-Weltmodell (TWM) als seine *Bedeutung* hat und als seine *Ausdrucksform* die sprachlichen Textstrukturen, die dieses schematische Strukturmuster-Modell im Sinne einer Genre-Grammatik kodieren (Schulze 2019: 15); vgl. die Bestimmung von Genres bei Schulze als allgemein handlungsbezogene kognitive Typisierungen, die sich im Besonderen auch auf Genres von Sprachhandlungen als Text-Genres und deren Weltmodelle bezieht:

> Genres stellen also semiotische Wissenseinheiten dar, denen die beiden Ebenen Weltmodell auf der kognitiven Seite und (strukturierte) Menge von Akten auf der Ausdrucksseite zugeordnet sind. Eine semiotisch spezifische Form des Genres ist zunächst durch einen spezifischen Typ von Akten charakterisiert, wobei im Folgenden rhetorische Akte betrachtet werden sollen. Weltmodelle, die mit einer mehr oder minder routinierten, strukturierten Sequenz von rhetorischen Akten gekoppelt sind, sollen im Folgenden als Text-Weltmodelle (TWM) bezeichnet werden [...]. (Schulze 2019: 8)

1.1.3 Quantitative Text-Weltmodell-Parameter

Die in TWM vorgenommene Typisierung der Struktur von Sprachhandlungssituationen basiert nach Schulze (2019) sowohl auf der schematischen Struktur des durch den Text als spezifische Abfolge von Sprachhandlungen beschriebenen Situationsmodells (ihrer Textur, s. Schulze 2019: 6) als auch auf den äußeren Bedingungen der Kommunikationssituation, d. h. auf dem nichtsprachlichen kommunikativen Kon-

text (s. Schulze 2019: 13),[10] dazu gehören etwa Sprecher, Hörer(schaft) sowie von beiden geteiltes Wissen.[11]

In dieser Arbeit, die sich mit Text-Weltmodellen nicht-dialogischer, insbesondere narrativer Texte beschäftigt, ist primär die Text-Modell-interne Strukturierung als Genre-konstituierendes Kriterium relevant, deren schematische Eigenschaften sich direkt in der sprachlichen Textstruktur widerspiegeln und entsprechend operationalisierbar und induktiv typisierbar sind. Im Rahmen einer vergleichenden Datenauswertung über Klassifikationsmethoden wird allerdings auch die Wechselwirkung zwischen den textinternen TWM-Strukturparametern und diskursfunktionalen Apriori-Kategorisierungen im Sinne textexterner Genre-Parameter untersucht, die sich also auf kontextuelle Struktureigenschaften der Kommunikationssituation beziehen (s. 5.2.4).

Bzgl. der textinternen Genre-Kriterien lassen sich neben den strukturellen auch substantielle sprachliche Genre-Marker feststellen, so z. B. die Einleitungsformel von Märchen, bei deren Verarbeitung automatisch das damit assoziierte Märchen-Text-Weltmodell aktiviert wird.[12] Primär drücken sich die in TWM gespeicherten, schematischen Genre-Regeln aber über strukturelle sprachliche Muster aus (vgl. Schulze 2019: 22, 31). Dies sind einerseits makro- bzw. mesostrukturelle Textmuster, die von satzübergreifenden Einheiten gebildet werden (etwa Dialogstrukturen, vgl. Schulze 2019: 16); vor allem aber sind es textuelle **Strukturmuster** auf der Mikro-Ebene, die von lexikalischen und grammatischen Einheiten verschiedener Komplexität gebildet werden.

Diese Textstrukturen werden über gebrauchsfrequenzbezogene kognitive Routinisierung (*entrenchment*, s. Langacker 2000: 3) typisiert, indem die in einem Kommunikationssituationstyp wiederholt auftretenden, rekurrenten Strukturen als schematische (typische) Muster abgespeichert werden: „The occurrence of psychological events leaves some kind of trace that facilitates their reoccurrence. Through repetition, even a highly complex event can coalesce into a well-rehearsed routine that is easily elicited and reliably executed“ (Langacker 2000: 3). Solche gebrauchsbasierten

10 Vgl. *context model* bei van Dijk (s. 2016: 67; 2018: 30; s. auch Ungerer & Schmid 1996: 45ff.).

11 Vgl. Schulze 2019: 10 (Hervorhebung im Original): „Wie oben schon angedeutet, haben TWM eine mehr oder minder elaborierte interne Struktur, die durch verschiedene Faktoren motiviert ist. Hierzu zählt primär natürlich die Semantik des jeweiligen rhetorischen Genres, dann aber auch Spezifika der Texttradition (wenn gegeben), der Redesituation, der Funktionalität, der Produzenten, des Adressatenkreis, der Intertextualität sowie der diastraten ebenso wie diaphasischen und auch diatopen Zuordnung. Hinzu treten das autosemantische Moment des Textes und vor allem die textinterne sprachliche Struktur. Grundsätzlich steht zu erwarten, dass das *Token* eines bestimmten rhetorischen Genres die meisten dieser Parameter abbildet […].“

12 So symbolisiert und aktiviert etwa der Standard-Marker deutscher Märchen *Es war einmal* das entsprechende Märchen-TWM. Ein Weltmodell kann also – z. B. durch einen solchen Genre-Marker – auch ohne gegebenen situativen Kontext, allein durch die Semantik oder die Struktur der textuellen Einheiten, aktiviert werden (s. Schulze 2019: 13).

quantitativen Texttypisierungen (*usage-based models*, vgl. Langacker 2000) sind nach Biber „[…] further defined quantitatively such that the texts in a type all share frequent use of the same set of co-occurring linguistic features. Because co-occurrence reflects shared function, the resulting types are coherent in their linguistic form and communicative functions“ (Biber 1992a: 332).

Basierend auf der frequenzbezogenen sowie der sequentiellen Verteilung linguistischer Einheiten in Texten bestimmt die Kognition in der **TWM-Schemainduktion** durch induktive Mustererkennung sukzessive die prototypischen Werte schematischer Eigenschaften von Texten eines Genres, indem die Vorkommen der von diesen Einheiten gebildeten textstrukturellen Musterregeln in verschiedenen Texten eines Genres – etwa über Durchschnittsbildung – quantifiziert werden:

> […] we can assume that frequency plays a crucial role with respect to the question, which domains are more strongly entrenched and hence more "grammatical" than others. The pragmasyntactic organization of an utterance heavily depends from whether or not both the speaker and the hearer are "used" to its information structure or not. In order to reconstruct the linguistic knowledge of a speaker as documented in text production, it hence seems reasonable to refer to the most frequent structural types. (Schulze 2004a: 552f.)

In der von der Kognition durchgeführten Induktion von Text-Weltmodellen wird also – basierend auf dem in der Sprachpraxis gegebenen Input von Textdaten – ein **quantitatives kognitives Modell** des regelmäßigen (schematischen) Auftretens bestimmter sprachlicher Einheiten aufgebaut, vgl. Croft 2016: 597: „A speaker's knowledge about her language is usage-based: it is a probability distribution of forms over meanings in the conceptual space, inferred from past usage events and constantly changing as forms are replicated to verbalize new experiences in new usage events.“

Diese kognitive Modellkonstruktion ist vergleichbar mit dem Aufbau statistischer Sprachmodelle in der quantitativen Computerlinguistik, so z. B. in der *grammar induction* (vgl. Anderson 1975; Pinker 1979), in der die Wahrscheinlichkeiten von Satzregeln aus deren Vorkommen in einem syntaktisch annotierten Korpus gelernt werden (s. Manning & Schütze 1999: 381ff.); vgl. auch Manning & Schütze 1999 zur Modellierung von Sprache und Kognition als probabilistische Phänomene:

> […] the cognitive processes used for language are identical or at least very similar to those used for processing other forms of sensory input and other forms of knowledge. These cognitive processes are best formalized as probabilistic processes or at least by means of some quantitative framework that can handle uncertainty and incomplete information. (Manning & Schütze 1999: 15)

So kann man z. B. annehmen, dass über die durchschnittliche Länge von Texten eines Genres (bzw. über die Frequenzverteilung der Textlängen) der prototypische Informationsgehalt für Text-Modelle dieses Genres repräsentiert ist. Texte des Genres „Slogan" haben beispielsweise einen in diesem Sinne stark komprimierten Informationsgehalt – repräsentiert durch ihre geringe Textlänge (vgl. Schulze 2019: 15). Ebenso kann man annehmen, dass über die Art und durchschnittliche Stärke von Tempus- und Aspekt-Markern im Text verschiedene kognitiv-schematische Zeitstrukturtypen repräsentiert sind. So haben verbale Einheiten in Texten des Genres „Kochrezept" im Deutschen typischerweise keine Tempusmarkierung, zeigen stattdessen aber durchgehend Imperativ- bzw. Infinitivformen, was man kognitiv als Repräsentation einer zeitlich nicht verorteten, stattdessen verhaltensbezogen-logischen Abfolgestruktur interpretieren kann.

Als die quantitativen Parameter einer kognitiven Texttypologie bilden solche, für ein Genre typische, signifikante Merkmale quantitativer Strukturmuster als Spur der der Textverarbeitung zugrunde liegenden mentalen Prozesse (vgl. Langacker 2000: 3; Schwarz 2000: 27; Schwarz-Friesel & Consten 2014: 11) die schematischen Modell-Parameter des dem Genre zugeordneten Text-Weltmodells ab. Diese **TWM-Parameter** repräsentieren demnach in ihrer Gesamtheit (vgl. Schulze 2020: 593) die statistischen Sprachhandlungsnormen einer Genre-Grammatik im Sinne einer genrespezifischen quantitativen Textgrammatik (van Dijk 1978; van Dijk & Kintsch 1978):

> Die Gesamtheit aller Einzelstimmen (oder Einzelpartituren), die für einen Text beschrieben werden können, zeigen an, wie ein bestimmtes TWM in einem Text-Token kodiert ist. […] Sicherlich werden keine zwei Texte über dieselbe Gesamtpartitur operieren. Aber es ist sehr wahrscheinlich, dass alle Texte eines bestimmten rhetorischen Genres durch prototypische Muster charakterisiert sind, die den *Type* der entsprechenden Gesamtpartitur repräsentieren. (Schulze 2019: 15, Hervorhebung im Original)

1.2 Forschungsvorhaben

1.2.1 Ziel der Arbeit

Ziel dieser Arbeit ist es, analog zur Abschätzung der quantitativ-schematischen Text-Weltmodell-Parameter (*parameter estimation*) durch das kognitive System des Menschen (vgl. Starfield 2005) in der **TWM-Mustererkennung**, computergestützte statistische Methoden zur Mustererkennung und Klassifizierung aus den Bereichen des maschinellen Lernens und der quantitativen Linguistik auf textstrukturelle Daten anzuwenden, um eine induktive Ableitung der gebrauchsbasierten Grammatik von

Texten eines Genres aus deren sprachlicher Struktur zu erreichen, die – indirekt[13] – auch das durch diese Menge von prototypischen Werten textstruktureller quantitativer Parameter kodierte kognitive Text-Weltmodell rekonstruieren lässt (s. Schulze 2019: 12); vgl. dazu auch Manning & Schütze 1999: 527: „Clustering can also be viewed as a form of category induction in cognitive modeling."

Vor dem theoretischen Hintergrund von sich gegenseitig ergänzender kognitiver und sozialer Erklärung der menschlichen Sprachtätigkeit über das in der Sprachpraxis stattfindende Erlernen von textuellen Informationsverarbeitungsregeln im Sinne einer *social cognition*[14] verbindet die Arbeit Methoden des **maschinellen Lernens** mit Parametern einer kognitiven **Textlinguistik**; dabei orientiert sie sich insbesondere an dem Ansatz zur datengestützten Rekonstruktion von Text-Weltmodellen über quantitative Textstruktur-Parameter von Schulze (2004a; 2018; 2019; 2020).[15] Während dort allerdings eine vergleichende Analyse der Textstruktur einzelner Volkserzählungen des Udischen als Grundlage für eine Herausarbeitung des darin kodierten Text-Weltmodells der Erzähltradition dieser kaukasischen Sprechergemeinschaft dient,[16] soll in dieser Arbeit eine entsprechende Auswertung mit automatischen Extraktions- und Analysemethoden auf einer größeren Anzahl von Texten durchgeführt werden. Damit strebt diese Arbeit die Umsetzung des in Schulzes Forschungsprogramm einer **TWM-Rekonstruktion** (2004a; 2018; 2019; 2020) angelegten Vorhabens eines automatisierten datengestützten Auffindens solcher gebrauchsbasierten kognitiven Texttypen über Textstrukturmuster (Heinemann & Viehweger 1991: 170, 174f.; Fix 2008: 66f., 71) in annotierten Korpora an:

> […] the diagnostic tools applied to the Udi data can be used at a larger scale to reconstruct the pragmasyntactic knowledge of speakers in a speech community. Naturally, in my paper I could apply these tools to selective data only. Many of the claims and tools

13 Als kognitives Konstrukt ist ein Text-Weltmodell nur im Resultat seiner Steuerung der Strukturierung in der Textproduktion beobachtbar.

14 Zum Begriff *social cognition* s. Augoustinos & Walker & Donaghue 2006: 7; Ochsner 2007: 42; vgl. auch van Dijk 2007: xxiv sowie van Dijk 2018: 31 zur Notwendigkeit der Verbindung von kognitiver und sozialer Psychologie für Diskursuntersuchungen.

15 Seine genrebezogene Theorie von Text-Weltmodellen hat Wolfgang Schulze, unter dessen Anregung und Anleitung diese Doktorarbeit ursprünglich entstanden ist, vor seinem unerwarteten Tod im Jahr 2020 vor allem in dem Buch-Manuskript Schulze 2018 („Schemas, Models, and Constructions. The Linguistic Representation of Event Image Structures in Grammar and Narration") ausgearbeitet. Daneben ist vor allem der Text Schulze 2019 bzw. 2020 zentrale Literatur für die theoretischen und methodischen Grundlagen dieser Arbeit: Schulze 2019 („Genre-gesteuerte linguistische Praktiken. Text-Weltmodelle von Volkserzählungen") ist das mir von Wolfgang Schulze zur Verfügung gestellte Artikel-Manuskript zu dem Ende 2020 in modifizierter Fassung unter dem Titel „Explorationen zur Genre-Grammatik von Volksnarrationen" posthum erschienenen Artikel Schulze 2020. Da eine komprimierte Darstellung der TWM-Theorie aus Schulze 2018 nur in dem Manuskript Schulze 2019 enthalten ist, nicht aber in dem veröffentlichten Text Schulze 2020, wird in dieser Arbeit auch auf dieses Artikel-Manuskript als Literatur zurückgegriffen.

16 Ähnlich auch die im selben Forschungskontext angesiedelte Studie Peng 2020 für Texte eines narrativen Subgenres im Quechua.

> have to be refined once they are applied to massive corpora. Nevertheless, I hope to have demonstrated that Cognitive Typology has not necessarily to end up in a mere perspective that would start with just a few data and would build a huge "house of interpretation" around these data. (Schulze 2004a: 573)

1.2.2 Korpusdaten

Für die angestrebte korpusbasierte Rekonstruktion von TWM wird eine Textsammlung von Volkserzählungen (Volksmärchen) sowie weiterer Genres (Mythen, Personal Songs, persönliche Erzählungen, journalistische Berichte) verwendet, die aus einer Sprechergemeinschaft mit in der Mündlichkeit verankerter Texttradition stammen – in diesem Fall sind das Texte aus Feldforschungsdaten der in Nordwestsibirien, hauptsächlich an den Flüssen Ob und Irtyš östlich des Urals ansässigen Khanten und Mansen, deren eng miteinander verwandte Sprachen Khanty und Mansi gemeinsam die **obugrischen** Sprachen bilden und die (üblicherweise mit dem Ungarischen als ugrische Sprachgruppe zusammengefasst) einen Zweig der uralischen Sprachfamilie darstellen.

Texte des Genres **Volkserzählung** – also ursprünglich mündlich überlieferte narrative Texte, die einerseits aus mnemotechnischen Gründen mit stereotypen Formeln arbeiten (vgl. Rubin 1995: 300), gleichzeitig aber auch Variabilität zeigen (s. Panzer 2020: § 27, § 30; Andersen 1997: 173; vgl. auch Schulze 2020: 605f.; Schulze 2000a: 4) – sieht Schulze als für die TWM-Mustererkennung deshalb besonders geeignet an, da für solche in mündlicher Sprachpraxis verankerte Volksmärchen anzunehmen ist, dass sich in ihnen – durch das Fehlen eines Editionsprozesses, wie er etwa bei den von den Brüdern Grimm gesammelten Märchen festzustellen ist – „die entsprechende linguistische Praxis unmittelbar abbildet" (Schulze 2019: 16).

Auch bei den obugrischen Texten des hier ausgewählten Korpus, die in Feldforschung gewonnen wurden, kann man davon ausgehen, dass sie die narrative Sprachpraxis der obugrischen Sprechergemeinschaft(en) relativ unverfälscht widerspiegeln; diese stellen also (auch vom Umfang her) den Kern des in dieser Arbeit untersuchten Korpus dar.

Da die Texte des verwendeten obugrischen Korpus (s. Kapitel 5) außerdem in syntaktischer, semantischer und informationsstruktureller Annotation vorliegen, sind diese für die angestrebte Methodenexploration besonders geeignet, da solche **Annotationsdaten** für die Berechnung der relevanten textstrukturellen Parameter notwendig sind (s. Kapitel 3; vgl. Kessler & Nunberg & Schütze 1997: 3).

1.2.3 Methodik

Ausgehend von den linguistischen Annotationsdaten werden für die Texte des Korpus durch **Feature-Construction**-Methoden (vgl. Motoda & Liu 2002: 69) verschiedene,

zuvor bzgl. ihrer Relevanz für den Aufbau kognitiver Text-Modelle diskutierte, textstrukturelle Parameter berechnet, um eine computergestützte Operationalisierung quantitativer TWM-Merkmale zu erreichen. Diese TWM-Parameter beziehen sich dabei auf die schematischen Text-Modell-Strukturbereiche der referentiellen und der relationalen Strukturierung sowie der Informationsstrukturierung. Die verschiedenen Parameter werden dabei durch Feature- bzw. Sequenz-Extraktionsmethoden in geeignete textstrukturelle Repräsentationen überführt; zum einen in Merkmalsmengen (sog. Feature-Sets, s. 4.1.1), zum anderen in globale sowie lokale Sequenzrepräsentationen (s. Abschnitt 4.4).

Anschließend können die Texte des Korpus unter Verwendung von theoriegestützten Textsorteneinteilungen (s. 5.2.4) zunächst einer deskriptiven Feature-Analyse bzgl. ihrer textstrukturellen Parameterwerte unterzogen werden, indem eine gruppierte Bestimmung der Werteverteilungen innerhalb dieser verschiedenen Apriori-Genre-Klassen durchgeführt wird. Durch Anwendung von automatischen Klassifizierungsmethoden erfolgt dann in einem sekundären Quantifizierungsschritt eine datenbasierte Analyse der Ähnlichkeitsstrukturen in den über TWM-Parameter repräsentierten Textdaten.

Im **Clustering** (vgl. Jain & Murty & Flynn 1999) als Methode der explorativen Datenanalyse (s. Abschnitt 4.2) können durch Berechnung von Distanzen zwischen über textstrukturelle Merkmalsmengen oder Sequenzen repräsentierten Texten Gruppen von Texten mit ähnlichen Textstrukturmustern identifiziert und somit induktive Clustertypologien erstellt werden. Die unterschiedlichen, dabei im Korpus jeweils für die Textcluster festgestellten Typen von Textstrukturmustern können dann – gemäß der Grundannahme dieser Arbeit zu Text-Weltmodellen als sprachgebrauchsbasierten Textstrukturmodellen, die situativ in der Textproduktion aktiviert werden – als potentieller Ausdruck (als die Spur) solcher kognitiver Genre-Modelle mit jeweils spezifischen prototypischen quantitativen TWM-Parameterwerten verstanden und entsprechend untersucht werden.

Daneben kommen auch verschiedene **Klassifikationsmethoden** als strukturprüfende Verfahren der Feature-Analyse zum Einsatz (s. Abschnitt 4.3). Insbesondere können mit sog. Feature-Selection-Methoden (vgl. Motoda & Liu 2002: 67f.) für textstrukturelle Merkmalsmengen Feature-Importance-Werte berechnet werden, die Aussagen über die Wichtigkeit der einzelnen Merkmale für die Differenzierung der jeweiligen Genre-Klassen der Texte im Korpus ermöglichen. Diese Ergebnisse können anschließend sowohl untereinander als auch – im Sinne einer externen Clusterevaluation – mit den entsprechenden Daten der induktiv entdeckten Clustergruppierungen verglichen werden. Eine solche Feature-Evaluation erlaubt also eine Gewichtung der untersuchten TWM-Parameter (s. Abschnitt 6.6) bzgl. ihrer Diskrimina-

tionsstärke für die verschiedenen (apriorischen sowie induktiven) genrebezogenen Texttypologien.[17]

Um eine solche datengestützte kontrastive Typisierung zu ermöglichen, umfasst das hier untersuchte obugrische Korpus nicht nur Volksmärchen als klassische Volkserzählungen, sondern – neben Texten aus mehr oder weniger funktional verwandten Genres wie mythologischen Sagen oder persönlichen Erzählungen und Berichten – auch einige Zeitungsberichte und Personal Songs, d. h. Vertreter von Textsorten, von denen man annehmen kann, dass sie eine abweichende textuelle Strukturierung aufweisen. Außerdem ist die Textauswahl zeitlich und räumlich-dialektal geschichtet, um eventuelle (z. B. zeitlich oder areal begründete) Cluster-Gruppierungen entdecken zu können; so kann sich ggf. zeigen, dass es in einer Sprechergemeinschaft nicht die *eine* narrative (usw.) Sprachpraxis gibt bzw. gab, sondern – zeitlich, areal oder durch den sozialen Kontext bedingt – unterschiedliche Strukturierungstypen von Erzähltexten vorkommen können. Andererseits können auch verschiedene Sprechergemeinschaften ein Text-Weltmodell teilen bzw. verwandte Modelle besitzen (eine gemeinsame Erzähltradition haben); dies gilt etwa für die Meddah-Tradition (vgl. Schulze 2019: 18), eine ursprünglich türkische Praxis des öffentlichen Geschichtenerzählens, die sich im gesamten Orient verbreitet hat (s. UNESCO 2021).

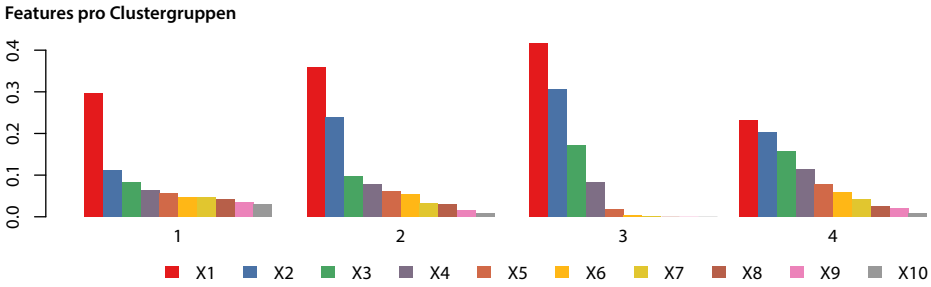
Als reines **Erprobungskorpus** für die in dieser Arbeit vorgestellten Auswertungsmethoden und Parametrisierungen ist das Korpus auf eine Auswahl von 34 Texten begrenzt, deren Umfang eine Interpretation der Ergebnisse unter Berücksichtigung der einzelnen Texte noch zulässt (s. Kapitel 5). Als Methodenexplorationsstudie im Sinne eines Pretests wird so die Verbindung hergestellt zwischen der auf der Analyse von Einzeltexten basierenden linguistischen Grundlagenforschung zu einer quantitativen kognitiven Texttypologie bei Schulze (2004a; 2018; 2019; 2020) und einer Anwendung der hier entwickelten und getesteten Verfahren und Methoden einer automatischen TWM-Mustererkennung auf großen Korpora zum Zweck einer automatischen Genre-Detektion.

Für dieses korpusbasierte Vorgehen sind im Rahmen einer automatisierten Feature-Construction Anpassungen der bei Schulze entwickelten Methoden und Parameter notwendig, auf die sich diese Arbeit stützt, vgl. Schulze 2004a: 573: „Many of the claims and tools have to be refined once they are applied to massive corpora."

So werden in dieser Arbeit beispielsweise statt einer manuellen, qualitativen Feststellung der Protagonisten bzw. Antagonisten (vgl. Schulze 2019: 23) die Hauptreferenten eines Textes im Rahmen einer automatischen Feature-Construction über die relative Häufigkeit ihrer Erwähnungen im Text bestimmt (s. 6.3.5). Auf den Daten dieser textweiten Frequenzverteilungen wird anschließend ein Clustering durchge-

17 Für globale Sequenzrepräsentationen kommen anstelle von Feature-basierten Klassifikationsmethoden spezielle Sequenz-bezogene Methoden zum Einsatz (s. 4.4.3).

führt, das die Texte des obugrischen Korpus in vier bzgl. der Topikalitätsstärke ihrer Hauptreferenten differenzierte Gruppen trennt (s. Plot 1.2.1). Die durchschnittlichen Frequenzverteilungen dieser Cluster können dann im Rahmen der TWM-Theorie als prototypische Werte der Referentenstruktur von dadurch potentiell mitkonstituierten Genres bzw. Subgenres interpretiert werden.



Plot 1.2.1: Clustergruppierte relative Textfrequenzen der häufigsten Referenten (Topikalitätsquotienten)

1.2.4 Forschungshorizont

Zusammenfassend ist das Ziel dieses Vorhabens die Erprobung von automatischen computerlinguistischen Methoden und Parametrisierungen für eine korpusbasierte Erstellung einer gebrauchsbasierten, induktiv gewonnenen, quantitativen kognitiven **Texttypologie** anhand einer Textsammlung obugrischer Volkserzählungen sowie Texten weiterer Genres. Das spezielle Forschungsinteresse dieser Arbeit besteht dabei in der Verbindung des Gegenstandsbereichs der kognitiven **Textlinguistik**, d. h. der Konstruktionsprozesse textbezogener kognitiver Modelle, mit Methoden des **maschinellen Lernens**, die beide auf der Klassifizierung von Weltinput zum Aufbau gebrauchs- bzw. allgemeiner datenbasierter Modelle beruhen.

So dienen die in dieser Arbeit eingesetzten quantitativen statistischen Methoden nicht nur als Hilfsmittel zur Beantwortung kognitiv-linguistischer Fragestellungen, wie dies im Rahmen hypothesenprüfender Analysen der Fall ist (vgl. etwa Frank 2019 als korpusbasierte Studie in einem Text-Welt-Theorierahmen): In Analogie gesetzt zur kognitiven Sprachverarbeitung (**Computermetapher**, s. Gaschler 2020: 366), können die formalen Klassifizierungsmodelle der maschinellen Sprachverarbeitung die Typisierung von rekurrenten Mustern des Sprachgebrauchs als Textstrukturregeln einer als Text-Weltmodell im Gedächtnis abgespeicherten Genre-Grammatik durch die Kognition **simulieren** (vgl. Abschnitt 2.3; Pinker 1979; Schütze 1997). Sie erlauben somit also eine Modellierung der mit einer solchen quantitativen kognitiven Texttypologie verbundenen kognitiven Lernprozesse.

Dabei liegt in der Darstellung dieser Verfahren ein Schwerpunkt darauf, die formalen Grundlagen sowie die konkreten (hier textlinguistischen) Anwendungsmöglichkeiten von Repräsentations- und Analysemethoden des maschinellen Lernens und

der Mustererkennung in einer für Sprachwissenschaftler zugänglichen Darstellung zu präsentieren, um so auch das Interesse an der Anwendung automatischer Lern- und Klassifizierungsmethoden in der linguistischen Forschung zu verstärken.

1.3 Kapitelübersicht

In Kapitel 1 werden das Forschungsvorhaben dieser Arbeit sowie dessen theoretische Grundannahmen vorgestellt. In Kapitel 2 erfolgt ein Überblick über den theoretischen und methodischen Hintergrund sowie den Forschungsstand zu diesem Vorhaben der Erstellung einer quantitativen kognitiven Texttypologie mit automatischen Klassifizierungs- und Mustererkennungsmethoden. Dabei werden die relevanten Disziplinen erörtert: Sprach- und Texttypologie, kognitive Textlinguistik, quantitative Linguistik und Korpuslinguistik sowie der computerlinguistische Aufgabenbereich der automatischen Dokumentenklassifizierung und des Text-Minings.

Im Anschluss daran gliedert sich die Arbeit in zwei Teile: Zunächst erfolgt mit Kapitel 3 und Kapitel 4 eine Methodendiskussion bzgl. der Operationalisierung und Erstellung einer quantitativen kognitiven Texttypologie. Kapitel 3 diskutiert verschiedene quantitative textstrukturelle Genre-Parameter, wie sie sich aus der Struktur lexikalischer, grammatikalischer, semantischer sowie pragmatisch-informationsstruktureller linguistischer Einheiten ergeben. Kapitel 4 stellt Methoden zur Repräsentation solcher Genre-Parameter vor sowie auch Methoden zur Klassifizierung von Texten anhand dieser Parameter, mit dem Ziel der Erkennung der ein Genre auszeichnenden Textstrukturmuster.

Daran anschließend erfolgt mit Kapitel 5 und Kapitel 6 eine Fallstudie (Pretest), in deren Rahmen die zuvor etablierten Methoden und Parameter am Beispiel eines syntaktisch und informationsstrukturell annotierten, geschichteten obugrischen Korpus primär narrativer Texte erprobt werden; nach einer soziokulturellen und sprachtypologischen Kurzeinführung in die obugrischen Sprachen werden in Kapitel 5 außerdem das Korpus samt Annotationsschema bzgl. der zuvor etablierten Grundparameter sowie die Annotationsparameter vorgestellt. Anschließend folgt in Kapitel 6 die Präsentation und Diskussion der Ergebnisse der Fallstudie, in der die zuvor besprochenen Methoden zur Operationalisierung und Erstellung einer quantitativen kognitiven Texttypologie auf das annotierte Korpus obugrischer Texte angewendet wurden; dazu gehört insbesondere auch eine Gewichtung der TWM-Parameter, eine Bewertung der Ergebnisse der Arbeit für die Methodendiskussion sowie eine Gesamtbewertung der Einzelergebnisse. Kapitel 7 schließt die Arbeit ab mit einem Fazit zur Generalisierbarkeit der in der Arbeit vorgeschlagenen Parametrisierung und Methodik für das übergeordnete Vorhaben einer Rekonstruktion von Text-Weltmodellen durch die Anwendung von automatischen Genre-Klassifizierungsmethoden.

2 Theoretischer und methodischer Hintergrund

2.1 Theoretischer Hintergrund

2.1.1 Texttypologie

Die Fragestellung einer gebrauchsbasierten funktionalen Typisierung von Texten kann allgemein einem linguistisch-typologischen Theorieansatz zugeordnet werden; anders als in sprachtypologischen Untersuchungen im engeren Sinne (vgl. van der Auwera & Nuyts 2007: 1075) wird in dieser Arbeit aber kein sprachübergreifender Vergleich angestrebt, sondern die kognitiv begründete Typisierung von Varianz in der Textproduktion einer Sprechergemeinschaft.[1] Während in den klassischen sprachvergleichenden Untersuchungen der Sprachtypologie die Variation struktureller Muster zwischen verschiedenen Sprachen Untersuchungsgegenstand ist, zielt die hier angestrebte Texttypologie stattdessen auf eine Typisierung des Sprachgebrauchs innerhalb einer Sprechergemeinschaft ab – diese Typologie bezieht sich damit auf die Variation sprachstruktureller Muster in Abhängigkeit von Gebrauchssituation und Kontext, deren Analyse entsprechend eine funktional-erklärende Untersuchung kognitiver Texttypen ermöglicht.

Gemäß der klassischen Texttypologie Isenbergs (1978; 1983) ist ein Texttyp als eine durch eine Menge an theoretisch begründeten Eigenschaften charakterisierte Textklasse zu verstehen, eine solche Theorie entsprechend als **Texttypologie** (vgl. Isenberg 1978: 566). Methodisch lassen sich zwei Arten von Texttypologien unterscheiden (vgl. Fix 2008: 70f.; Gautier 2009: 1f.): Apriori-Typologien, deren Klassen schon im Voraus (*top-down*) nach bestimmten theoretischen Überlegungen als feststehende Größen festgelegt sind (s. 5.2.4), und induktive Texttypologien, deren klassenspezifische texttypologische Eigenschaften *bottom-up* aus empirischen Daten gewonnen werden (s. Fix 2008: 69f.; vgl. Heinemann & Viehweger 1991: 114).[2]

In dieser Arbeit wird primär eine Texttypologie im Sinne des induktiven Ansatzes angestrebt, indem Texte ausgehend von den in ihnen vorfindlichen Mustern klassifiziert werden, wodurch eine gebrauchsbasierte Feststellung von Texttypen anhand

1 Vgl. Croft 2016: 589: „In order to make the connection between typology and Cognitive Linguistics, the cross-linguistic patterns of linguistic diversity – Greenbergian universals – must somehow emerge from the behavior of individual speakers in each speech community. Since Greenbergian universals are patterns of diversity, that is, of cross-linguistic variation, an obvious place to look for the connection is language-internal variation."

2 Diese Einteilung der Texttypologisierung bzw. Textklassifizierung korrespondiert methodisch mit der Differenzierung von überwachten computerlinguistischen Lernmethoden (Textklassifikation durch Zuordnung zu gegebenen Gruppen) und unüberwachten Lernmethoden (induktive Typisierung durch Textclustering); s. 2.2.2.

der empirisch vorfindlichen **Textmuster** (Fix 2008: 67; s. auch Gautier 2009: 1f.) ohne feste Zuordnung von Texten in Klassen nach Apriori-Kriterien wie in der klassischen Texttypologie von Isenberg ermöglicht wird (s. Fix 2008: 67ff.); theoriebasierte Apriori-Einteilungen (etwa nach klassischen Textsorten) werden in dieser Arbeit nur ergänzend im Rahmen der Analyse der induktiven Clustertypisierungen für einen Vergleich mit den dort empirisch in den Daten festgestellten Strukturmustertypen verwendet.

Für die beabsichtigte empirisch-induktive Bestimmung von Texttypen im Sinne von Textstrukturmustern durch statistische Methoden agiert eine hier angestrebte gebrauchsbasierte texttypologische Untersuchung mit quantitativen Parametern (vgl. Kapitel 1). Als theoretische Grundlage, auf der die hier angestrebte Feststellung von Texttypen basiert, dient gemäß der in Kapitel 1 getroffenen kognitiven Grundannahmen und dem daraus resultierenden Vorhaben, Text-Weltmodelle im Sinne von genrespezifischen Wissensstrukturen zu extrahieren, die kognitive Textlinguistik.

2.1.2 Kognitive Textlinguistik

Für eine strukturbezogene Texttypologie im Sinne einer Texttypen-Theorie (vgl. Isenberg 1978: 566) können in der Textlinguistik entwickelte, textstrukturelle Analysekriterien herangezogen werden: Zu diesen **Textualitätskriterien** (vgl. de Beaugrande & Dressler 1981) gehören insbesondere die Kohärenzstruktur von Texten als inhaltsbezogene Relationen zwischen den textkonstituierenden Einheiten (Phrasen, Sätze, Absätze), also vor allem referentielle und relationale Kohärenzbeziehungen (vgl. Sanders & Spooren 2007: 919), sowie Kohäsionsmittel als sprachlicher Ausdruck dieser Kohärenzrelationen (u. a. über Anaphorik, Wiederholung, Elliptik, Tempus).

Seit den 1970er-Jahren hat sich die Textlinguistik in der sog. **kognitiven Wende** (s. Figge 2000; Fix 2008: 68) kognitionspsychologischen Erklärungen für die Entstehung, die Funktion und den Gebrauch von textuellen Strukturen zugewandt; gemeinsam ist diesen Ansätzen, dass sie Texte als Sprachhandlungskomplexe sehen, mit denen „auf Objekte und Sachverhalte der außersprachlichen Welt referiert" (Schwarz-Friesel & Consten 2014: 73) wird, die also der Übertragung von referentiell und relational strukturierten Informationen dienen (s. Kapitel 1). Eine solche kognitive Textlinguistik bezieht sich also mit der textuellen Informationsverarbeitung durch die menschliche Kognition auf bedeutungs- und informationsstrukturelle Textualitätskriterien (Handlungs- und Referentenstruktur, neue vs. bekannte Information, Aufmerksamkeitssteuerung usw.).

Historisch beginnt die kognitive Textlinguistik mit den psychologisch-linguistischen Ansätzen zum Textverstehen von van Dijk und Kintsch (van Dijk 1972; Kintsch & van Dijk 1978), die sich mit textuellen Gedächtnisstrukturen beschäftigen (Makro- und Mikrostrukturen; später in gemeinsamer Arbeit ergänzt um die Kon-

zeption des Situationsmodells zur Erklärung der Anbindung von textueller Informationsverarbeitung an Weltwissen, s. van Dijk & Kintsch 1983; vgl. Kapitel 1). Vorläufer der kognitiven Textlinguistik bzgl. der Frage nach den Texten zugrunde liegenden Strukturen ist die literaturwissenschaftlich orientierte, funktionale Analyse der Bedeutungsstruktur narrativer Texte von Propp (1972; Erstveröffentlichung 1928), die sich auf einen den Texten zugrunde liegenden Handlungsstrukturtyp bezieht;[3] ebenso die daran anknüpfende Erzählstrukturanalyse von Lévi-Strauss (1972) zur semantischen Struktur von Mythen sowie die Analyse der Struktur von Ereignisabfolgen und das Aktantenmodell von Greimas (1972; vgl. Renner 2000) sowie die *story grammar* von Rumelhart (1975), die in Anschluss an Propp Handlungsstrukturregeln des narrativen Textaufbaus beschreibt. Eine weitere Ursprungstradition der kognitiven Textlinguistik kann im distributional-semantischen Ansatz von Harris (1952; 1954) gesehen werden (vgl. Figge 2000: 97ff.).

2.1.3 Genre als kognitive Texttypisierung

Ein wichtiges Feld in der aktuellen kognitiven Textlinguistik sind Text-Welt-Theorien (vgl. Kapitel 1), so etwa Werth (1999), Schwarz (2000, 2001, 2011) oder Gavins (2007). Im Zentrum dieser Theorien steht die Analyse des wissensbezogenen Aufbaus einer ‚Textwelt' durch Elaborations- und Inferenzstrategien – also der mentalen Repräsentation eines Textes als sog. „text-world model, which is not a text level but a mental level of referential structures" (Schwarz-Friesel & Consten 2011: 352).

Die Text-Weltmodell-Theorie von Schulze (2018; 2019; 2020), an die diese Arbeit anknüpft (vgl. Kapitel 1), bietet dagegen als sozio-kognitiv ausgerichteter Ansatz ein genrebezogenes Verständnis von kognitiven **Text-Weltmodellen**, nämlich als durch Typisierung der Struktur von Texten, die in ähnlichen Kommunikationssituationen auftreten, induktiv gelernte, schematische Strukturmodelle. Diese gebrauchsbasierten Typen rekurrenter quantitativer Strukturmerkmale von Textmodellen werden situativ aktiviert und steuern den Textmodell-Aufbau in der Textproduktion und -verarbeitung: Sie verkörpern also Musterregeln, wie man in einer bestimmten Situation spricht. Wie in 1.1.2 dargestellt, nennt man eine solche funktionale, sprachhandlungsbezogene Texttypisierung – von Swales definiert als „class of communicative events […] [*with*] some shared set of communicative purposes" (Swales 1990: 45f.) – auch das **Genre** eines Textes (Miller 1984; Schulze 2019):

3 Vgl. Propp 1972: 9: „Indessen ist auf dem Gebiet des Volksmärchens eine Formanalyse sowie die Ableitung von Strukturgesetzmäßigkeiten ebenso gut möglich wie bei Organismen. Wenn diese Behauptung auch nicht für alles, was als Märchen bezeichnet wird, zutrifft so doch auf jeden Fall für die sogenannten Zaubermärchen […]."

> Over the past thirty years, researchers working across a range of disciplines and contexts have revolutionized the way we think of genre, challenging the idea that genres are simple categorizations of text types and offering instead an understanding of genre that connects kinds of texts to kinds of social actions. As a result, genres have become increasingly defined as ways of recognizing, responding to, acting meaningfully and consequentially within, and helping to reproduce recurrent situations [...] as typified rhetorical ways of interacting within recurring situations [...]. (Bawarshi & Reiff 2010: 3)

Im Rahmen der in dieser Arbeit angestrebten quantitativen kognitiven Texttypologie werden dementsprechend Methoden für die datenbasierte, automatische Klassifizierung von Genres als solchen gebrauchsbasierten (d. h. kognitiv begründeten und sozial vermittelten) Texttypen erprobt. Für die Operationalisierung der Parameter dieser kognitiven Texttypologie in Kapitel 3, d. h. für die Untersuchung der mentalen TWM-Strukturmodelle, auf denen die genrespezifische Struktur von Texten basiert, bedient sich die Arbeit dabei verschiedener, in der kognitiven Textlinguistik entwickelter Textualitätskriterien.

2.2 Methodischer Hintergrund

2.2.1 Quantitative Textlinguistik

Für die Operationalisierung der strukturellen TWM-Modell-Parameter sind insbesondere Methoden der **quantitativen Textlinguistik** relevant; dieses auf Textstatistik basierende Teilgebiet der quantitativen Linguistik beschäftigt sich mit der „statistischen Textorganisation" (Mehler 2005: 325f.) und betont damit „die Bedeutsamkeit quantitativer Aspekte von Textproduktion und -rezeption" (Mehler 2005: 326; s. auch Altmann & Gerlach 2016):

> Im Kern beruht der Textbegriff der quantitativen Linguistik auf der Auffassung, daß Texte quantitative Eigenschaften und Relationen aufweisen, die für ihre Struktur und Organisation konstitutiv sind. Textualität wird als eine Eigenschaft verstanden, die u. a. an der spezifischen statistischen Organisation textueller Einheiten zum Ausdruck kommt. Die Verteilungen quantitativer Texteigenschaften bilden sozusagen textspezifische Ordnungszustände. (Mehler 2005: 325)

Ein zentraler Bereich der quantitativen Textlinguistik sind texttypologische Analysen, die auch unter dem Begriff der quantitativen **Stilistik** gefasst werden (vgl. Mehler 2005: 325, 341f.). In solchen textstatistisch-typologischen Untersuchungen (s. z. B. bei Grzybek & Kelih & Stadlober 2005) werden verschiedene quantitative textstrukturelle Parameter zur Untersuchung von Texttypen verwendet, die auf „Texteinheit, Texteigenschaft und Wiederholung (Rekurrenz) sowie [...] [*den*] hieraus ableitbaren Begriffe[*n*] der Häufigkeit (Frequenz), Verteilung (Distribution), Länge, Kookkur-

renz, Textposition und des Textsegments“ (Mehler 2005: 330) basieren. Dazu gehören zunächst einfache Indizes wie Type-Token-Verhältnisse (d. h. Maße textueller Varianz über Wortwiederholung), Wort- und Satzlängenmaße oder Content-Word-bezogene Maße des lexikalischen Reichtums (vgl. Mehler 2005: 333, 340; s. auch Virtanen 2009: 1073ff.). Aber auch quantitative Untersuchungen der textuellen Referenzstruktur über die Verteilung phorischer Ausdrücke sowie die Untersuchung der syntagmatischen Topikalitätsstruktur über quantitative Parameter wie die referentielle Distanz (Givón 1983b) sind Beispiele einer quantitativen Textlinguistik (vgl. Mehler 2005: 338, 342).

Als grundsätzlich empirisch-gebrauchsbasiert orientierte Disziplin bedient sich die quantitative Textlinguistik in diesen Analysen primär korpuslinguistischer Daten, also Textsammlungen als repräsentative Samples eines spezifischen Sprachgebrauchs, auf denen über quantitative Parameter generalisierbare Sprachgebrauchsmuster beschrieben und ausgewertet werden können (s. Mehler 2005: 326; Biber & Jones 2009: 1287; vgl. auch Biber & Conrad & Reppen 1998; Bubenhofer 2008). Solche **korpusbasierten** textlinguistischen Untersuchungen sind als Analysen der textspezifischen Verteilung verschiedener quantitativer Parameter, wie normierter Textfrequenzen oder Kookkurrenzen linguistischer Kategorien, ein Haupttyp korpuslinguistischer Analysen (Biber & Jones 2009: 1298f.).

Als Textklassifikationsmethoden kommen in der quantitativen Textlinguistik verschiedene Diskriminanz- und Varianzanalysen wie LDA, ANOVA oder GLM zum Einsatz (z. B. bei Biber 1992b; Grzybek & Kelih & Stadlober 2005; vgl. auch 4.3.4.2); diese Verfahren ermöglichen einen Vergleich der Fähigkeit quantitativer Parameter zur Differenzierung von vorgegebenen Textklassen. Ebenso finden in der quantitativen Textlinguistik auch Clusteranalysen Anwendung (so z. B. bei Grzybek & Kelih & Stadlober 2005), die durch die Erstellung von induktiven, datenbasierten Texttypisierungen eine explorative Untersuchung der in Textsammlungen auftretenden Strukturmuster erlauben. Dabei werden die Texte über das sog. Datenrepräsentationsmodell als Mengen (Feature-Sets) quantitativer Strukturmerkmale operationalisiert (Mehler 2005: 341; s. auch Abschnitt 4.1).

Von diesen Ansätzen wird in dieser Arbeit die Methodik der Repräsentation von Texten über Mengen quantitativer Parameter übernommen und mit Hilfe von kognitiv-textlinguistischen Parametern erprobt; dementsprechend werden Textklassifizierungsmethoden der quantitativen Textlinguistik, wie Diskriminanzanalyse und Clustering, zur Erstellung von quantitativen Texttypologien verwendet.

2.2.2 Automatische Dokumentenklassifizierung

Neben der stark forschungsbezogenen quantitativen Linguistik hat auch die quantitative Computerlinguistik als Anwendungswissenschaft (auch: NLP = *Natural Language Processing*) statistische Methoden zur Dokumentenklassifizierung entwickelt,

die dem Aufbau von Modellen zur automatischen Texttypisierung dienen (s. Manning & Schütze 1999: 575-608; vgl. auch Mehler 2005: 341). Die automatische Textklassifikation geht dabei üblicherweise von quantitativen lexikalischen Merkmalen aus (vgl. aber Abschnitt 2.3), liefert also im Gegensatz zur strukturellen Analyse in der quantitativen Linguistik eine themenbasierte Texttypisierung (Manning & Schütze 1999: 575; vgl. Kessler & Nunberg & Schütze 1997: 32).

In dieser Ausrichtung ist die automatische Textklassifizierung (Document-Categorization, Manning & Schütze 1999: 529-574) eng verwandt mit der Disziplin des Information-Retrieval, die Algorithmen und Methoden zur Extraktion und zum Auffinden von Informationen in Dokumentensammlungen entwickelt (Manning & Schütze 1999: 529; s. auch Manning & Raghavan & Schütze 2009), sowie mit dem Bereich des Text-Mining (Heyer & Quasthoff & Wittig 2008), das Mustererkennungsmethoden auf Textdaten anwendet, um explorativ Informationen aus den Textdokumenten zu gewinnen (vgl. Beyerer & Richter & Nagel 2017; Aggarwal 2015; 2018; Tan & Steinbach & Kumar 2006).

Die in diesen Bereichen eingesetzten Methoden und Algorithmen können allgemein dem Bereich des maschinellen Lernens (ML) zugeordnet werden; zum Einsatz kommen hier neben klassischen statistischen Lernmethoden (lineare sowie nichtlineare Modelle, z. B. SVM; s. dazu Kapitel 4; vgl. James u. a. 2017), die auch in der quantitativen Linguistik eingesetzt werden, verstärkt mehrschichtige neuronale Netze (Deep-Learning; so etwa bei Kessler & Nunberg & Schütze 1997 für die Genre-Klassifikation). Wie bei den Methoden der quantitativen Linguistik ist auch hier zu unterscheiden zwischen Methoden des überwachten Lernens (**Textklassifikation** als Lernen einer Funktion, die Merkmalrepräsentationen von Texten auf Apriori-Textklassen abbildet) und Methoden unüberwachten Lernens als datengestützte Induktion von Textklassen (insbesondere **Textclustering** als automatische Klassifizierung über die Feststellung von Gruppen mit ähnlichen Merkmalen).

Von diesen Ansätzen wird in dieser Arbeit das Vektorraummodell als die Repräsentation von Dokumenten als Vektoren multidimensionaler Merkmalsräume sowie die Berechnung von Ähnlichkeiten zwischen diesen Dokumentrepräsentationen für explorative Clusteringanalysen relevant sein. Außerdem kommen verschiedene Methoden und Algorithmen der statistischen Dokumentenklassifikation zum Einsatz, insbesondere Ensemble-Methoden für eine Feature-Selection; diese dient in der automatischen Dokumentenkategorisierung primär der Optimierung von Klassifikationsmodellen, wird in dieser Arbeit aber zur Analyse von quantitativen, kognitiv-textlinguistischen Feature-Sets eingesetzt.

2.3 Forschungsstand

Dieser Abschnitt gibt eine Übersicht über den Stand der Forschung in den oben besprochenen, für die korpusbasierte Erstellung einer quantitativen kognitiven Genre-Typologie in dieser Arbeit relevanten Disziplinen.

Formale Modellierung kognitiver Lernprozesse. Mit der Simulation des menschlichen Spracherwerbs durch Computermodelle beschäftigen sich in der kognitiven Psychologie und Linguistik u. a. Anderson 1975 (mit Überlegungen zu einer *grammar induction*) und Pinker 1979: „[…] it may be necessary to find out how language learning *could* work in order for the developmental data to tell us how it *does* work" (1979: 280, Hervorhebung im Original). Schütze 1997 verwendet mit Klassifikationsverfahren der Computerlinguistik statistische Machine-Learning-Modelle zur Induktion syntaktischer und semantischer Kategorien durch unüberwachte Klassifizierung mit agglomerativem Clustering bzw. zum überwachten Lernen von Subkategorisierungsframes mit neuronalen Netzen (s. auch Rumelhart & Hinton & Williams 1986). Schütze zeigt, dass solche aus unannotierten Korpora gelernte, statistisch-quantitative Modelle die hier relevanten Lernprozesse der menschlichen Kognition adäquat zu simulieren im Stande sind und dass sie zudem durch die graduelle Repräsentation der Kategorien gut mit Ambiguität umgehen können. Auch wenn sich diese Ansätze auf die Induktion von Sprachwissen beziehen, so liefern sie doch die Vorlage für das Vorgehen in dieser Arbeit zur Induktion von Textsortenwissen über die Erstellung von korpusbasierten quantitativen Modellen von gebrauchsbasierten Texttypen als Analysemodelle für potentielle analoge kognitive Text-Weltmodelle.

NLP-Ansätze zur automatischen Genre-Klassifikation. Die Aufgabe der automatischen Identifizierung von Genres als funktionale Texttypen wird in der quantitativen Computerlinguistik u. a. bei Kessler & Nunberg & Schütze 1997 diskutiert;[4] sie weisen dabei insbesondere auf die Schwierigkeiten hin, die mit der Operationalisierung einer solchen funktionalen Texttypisierung verbunden sind:

> Another reason for the neglect of genre, though, is that it can be a difficult notion to get a conceptual handle on, particularly in contrast with properties of structure or topicality, which for all their complications involve well-explored territory. In order to do systematic work on automatic genre classification, by contrast, we require the answers to some basic theoretical and methodological questions. Is genre a single property or attribute that

4 Vgl. Kessler & Nunberg & Schütze 1997: 33: „We will use the term 'genre' here to refer to any widely recognized class of texts defined by some common communicative purpose or other functional traits […]."

> can be neatly laid out in some hierarchical structure? Or are we really talking about a multidimensional space of properties that have little more in common than that they are more or less orthogonal to topicality? (Kessler & Nunberg & Schütze 1997: 32)

Karlgren & Cutting 1994 verwenden für eine Genre-Klassifikation durch Diskriminanzanalysen einfache textstatistische und strukturelle Features (Type-Token-Verhältnis, Frequenzen von Wortarten usw.). Kessler & Nunberg & Schütze 1997 überprüfen im Anschluss daran, ausgehend u. a. von strukturellen bzw. lexikalischen Merkmalen, die Grundlage für die entsprechenden struktur- bzw. themenbezogenen NLP-Klassifikationsaufgaben sind, welche Feature-Typen sich für die Klassifikation von Textgenres als funktionale Texttypen eignen:

> Examples of structural cues are passives, nominalizations, topicalized sentences, and counts of the frequency of syntactic categories (e. g., part-of-speech tags). […] Most facets are correlated with lexical cues. Examples of ones that we use are terms of address (e. g., *Mr.*, *Ms.*) […]. Character-level cues are mainly punctuation cues and other separators and delimiters used to mark text categories like phrases, clauses, and sentences […] Derivative cues are ratios and variation measures derived from measures of lexical and character-level features. (Kessler & Nunberg & Schütze 1997: 34)

So zeigen sie u. a., dass Oberflächenfeatures (Interpunktion etc.) eine ähnliche Stärke in der Genre-Diskrimination liefern wie strukturelle Merkmale.

Finn & Kushmerick 2006 sehen, ähnlich wie Karlgren & Cutting 1994, die Aufgabe der Genre-Klassifikation als mit der Stilistik verwandt:

> The genre of a document reflects a certain style rather than being related to the content. In general, this is what we mean when we refer to the genre of a document: The genre describes something about what kind of document it is rather than what topic the document is about. (Finn & Kushmerick 2006: 1507)

Finn & Kushmerick testen erfolgreich verschiedene Features (s. 2006: 1511f.) für die Textklassifizierung nach Genreklassen, nämlich Bag-of-Words-Modelle, Part-of-Speech-Muster und einfache Textstatistiken (das sind „document-level statistics […][:] sentence length, number of words, word length", Finn & Kushmerick 2006: 1512; ebenso Frequenzen von Funktionswörtern und Punktuation; vgl. auch Nguyen u. a. 2012: 378). Dabei stellen sie eine partielle Topikabhängigkeit ihrer Genre-Modelle fest: „In theory, genre and topic are orthogonal; however, our experiments indicate that in practice they partially overlap" (Finn & Kushmerick 2006: 1517).

Methoden der Genre-Klassifizierung in der quantitativen Linguistik. In mehreren korpuslinguistischen Studien (Biber 1992a; 1995; 2014) beschäftigt sich Biber mit multivariaten Untersuchungen von Registervariation über Faktor- und Cluster-

Analysen;[5] der Schwerpunkt liegt hier auf der Überprüfung von Hypothesen bzgl. der genrespezifischen Verwendung bestimmter Konstruktionen (vgl. Kessler & Nunberg & Schütze 1997: 33). Verwendet werden quantitative Daten zu strukturellen linguistischen Features wie Nominalisierung, TAM-Marker, Adverbiale oder Subordination (Biber 1992a: 333; vgl. Kessler & Nunberg & Schütze 1997: 34). Grzybek & Kelih & Stadlober 2005 verwenden für ihre auf verschiedenen Wortlängen-Maßen basierende „quantitative Texttypologie" Diskriminanz- und Clusteranalysen zur stilistischen Auswertung eines umfangreichen Korpus und stellen fest, dass „eine Gruppierung der Texte nach Funktionalstilen den Ergebnissen der quantitativen Untersuchungen nicht standhält. Stattdessen ergibt sich die Notwendigkeit und Möglichkeit, das gesamte Text-Universum in drei bzw. vier Diskurstypen zu untergliedern, deren Relevanz und Tragfähigkeit sowohl im Hinblick auf weitere Texte als auch auf andere Sprachen zu prüfen sein wird" (Grzybek & Kelih & Stadlober 2005: 118). In Legallois & Charnois & Larjavaara 2018b sind verschiedene disziplinübergreifende quantitative linguistische Ansätze versammelt, „in which genre analysis is an essential parameter for understanding" (Legallois & Charnois & Larjavaara 2018a: 1), also Ansätze, in denen Genre erklärende Funktion hat; für die Fragestellung hier sind insbesondere Schnedecker 2018 mit einem Ansatz zur Genre-Identifikation über Referenzketten sowie Lapshinova-Koltunski & Zampieri 2018 mit einer Feature-Analyse im Rahmen einer Klassifikationsaufgabe (Textklassifikation für Genre-Features) relevant. Die Studien in diesem Sammelband erproben zusätzlich zu den paradigmatischen Parametern klassischer quantitativer Textlinguistik (Legallois & Charnois & Larjavaara 2018a: 2), wie Type-Token-Relation, Satzlänge oder lexikalischer Reichtum, auch sequentielle Operationalisierungen. Solche (bei Legallois & Charnois & Larjavaara 2018a als ‚nicht-diskret' bezeichneten) Parametrisierungen für textstrukturelle Repräsentationen – wie sie etwa bei Lapshinova-Koltunski & Zampieri 2018 für die genrebezogene Feature-Analyse, basierend auf Part-of-Speech-N-Grammen mit statistischen Klassifikationsmethoden, angewendet werden – spielen auch in dieser Arbeit eine wichtige Rolle (vgl. Kapitel 3, Partitur-Folgen bei Schulze 2019 als zentraler TWM-Muster-Typ).

Kognitiv-informationsstrukturelle Features in der quantitativen Linguistik. Wie Biber & Conrad & Reppen feststellen (1998: 106f.; vgl. auch Gries & Stefanowitsch 2007), gibt es aufgrund der schwierigen Operationalisierbarkeit solcher pragmatisch-informationsstrukturellen Merkmale nur wenige korpuslinguistische Untersuchun-

5 Vgl. Biber 1992a: 332: „In the series of studies using the approach, I have used the term 'genre' (or 'register') for text varieties that are readily recognized and 'named' within a culture (e. g., letters, press editorials, sermons, conversation), while I have used the term 'text type' for varieties that are defined linguistically (rather than perceptually). Both genres and text types can be characterized by reference to co-occurring linguistic features."

gen, die pragmatisch-informationsstrukturelle Parameter und diskurstypologische Fragestellungen verbinden, sich also mit dem Zusammenhang des Gebrauchs von grammatischen Konstruktionen mit Diskurstypen beschäftigen, z. B. Turn-Taking-Strukturen in Konversationen, dem Referententracking oder der Verteilung von neuen vs. bekannten Informationen (s. Biber & Conrad & Reppen 1998: 106f.). Als eine der wenigen quantitativen korpuslinguistischen Untersuchungen bzgl. Diskursmuster-Typen verwendet Biber 1992b (vgl. auch Biber & Conrad 2001) referentiell-informationsstrukturelle quantitative Features, wie referentielle Distanz oder die Verteilung neu in den Diskurs eingeführter Informationseinheiten (referentielle Informationsdichte), zur korpuslinguistisch-statistischen Untersuchung der Genre-Distribution dieser referentiellen Features mit linearen Modellen.

Quantitative diskurstypologische Parameter in der Sprachtypologie. Auch in der mit quantitativen Parametern agierenden Sprachtypologie gibt es einige prominente Ansätze mit diskurspragmatischen Parametern wie etwa Metriken referentieller Distanz oder Dichte (u. a. Hopper & Thompson 1980; Givón 1983b; Cooreman 1987; Du Bois 1987; Bickel 2003; s. Myhill 2001 für einen Überblick), in denen der Einfluss pragmatischer Informationsstrukturierung auf die grammatikalische Sprachstruktur untersucht wird. Dazu gehören auch korpusgestützte diskurstypologische Untersuchungen, in denen hypothesenprüfend (vgl. Gries 2008: 294) der Zusammenhang grammatischer Variation mit Genres (als sprachinterne informationsstrukturelle Typen) untersucht wird (z. B. in Dorgeloh & Wanner 2010; vgl. Legallois & Charnois & Larjavaara 2018a: 1). Anders als mit der in dieser Arbeit geplanten, datenbasierten Induktion von kognitiven Texttypen im Sinne eines korpusbasierten Vorgehens (*corpus-driven*, vgl. Tognini-Bonelli 2001: 84f.) wird in diesen sprachtypologischen Ansätzen allerdings eine grammatikbezogene Analyse des Einflusses des Diskurstyps auf die Syntax durchgeführt, also eine Untersuchung von „genre effects on syntax" statt einer datengestützten Induktion von „grammar of genres" (Legallois & Charnois & Larjavaara 2018a: 1), wie sie in dieser Arbeit angestrebt wird.

Studien zur automatischen Klassifizierung von Volkserzählungen. Mit der Untersuchung der automatischen Klassifizierung von Volkserzählungen und ihrer Subgenres beschäftigen sich Nguyen u. a. 2012 (auch Meder u. a. 2016), die verschiedene Features zur SVM-Klassifikation eines niederländischen Korpus von Volkserzählungen nach Subgenres (Legenden, Märchen usw.) überprüfen – wobei sich Character-N-Gramme als bester der untersuchten Feature-Typen herausstellen (Nguyen u. a. 2012: 382). Reiter & Frank & Hellwig 2014 stellen am Beispiel eines Korpus von Volkserzählungen sowie eines Ritualtext-Korpus Klassifikations- und Clusteringmethoden zur Extraktion narrativer Bedeutungsstrukturmuster über semantische Diskursrepräsentationen vor; für diese narrativ-strukturellen Features verwenden sie dabei u. a. auch

sequentielle Repräsentations- und Alignment-Methoden, z. B. für häufige Ereignisabfolgen (s. Reiter & Frank & Hellwig 2014: 585).

Automatische Extraktion semantischer narrativer Strukturen. Ähnlich wie Reiter & Frank & Hellwig 2014 entwickeln auch Chambers & Jurafsky 2009 und Finlayson 2016 computerlinguistische Methoden zur automatischen Induktion von narrativen schematischen Mustern, nämlich Bedeutungsstruktur-Handlungsrahmen (relationale Frames und Partizipantenrollen) als Folgen von Ereignis-Slots (s. Chambers & Jurafsky 2009: 604) – bei Chambers & Jurafsky 2009 über Informationen zu Subkategorisierungsrahmen aus der FrameNet-Datenbank, bei Finlayson 2016 (der eine automatische Extraktion von Propps ‚Funktionen' anstrebt, also der dort für die russischen Zaubermärchen festgestellten Handlungsstrukturmuster, vgl. Propp 1972) über Frame-Informationen aus der PropBank-Datenbank, basierend auf den englischen Übersetzungen der Zaubermärchen.

Quantitative kognitive Merkmale von Volkserzählungen. Im Rahmen seiner Text-Weltmodell-Theorie (s. 2.1.3) entwickelt Schulze (2004a; 2018; 2019; 2020) eine gemäß verschiedener kognitiver Domänen (Makrostrukturbereiche, s. Schulze 2020: 606f.) geordnete Parameter-Systematik von quantitativ operationalisierbaren Strukturmerkmalen für eine Operationalisierung kognitiver Genre-Modelle von Volkserzählungen, die in dieser Arbeit unter Verwendung von Methoden des Data-Minings sowie automatischer Klassifizierungsmethoden erprobt werden soll (s. Abschnitt 1.2). Neben einfachen textstatistischen Merkmalen (Type-Token-Verhältnissen usw.) und frequenzbezogenen textstrukturellen Features (Referentenhäufigkeit usw.) sind hier mit dem Konzept der Partitur-Folge insbesondere auch sequentielle Textverteilungen quantitativer Variablen zentral.

3 Parameter einer quantitativen kognitiven Texttypologie

Kapitelzusammenfassung

In diesem Kapitel werden quantitative Textstruktur-Parameter diskutiert, die gemäß der in Kapitel 1 getroffenen Grundannahmen geeignet sind zur datengestützten Mustererkennung für die korpusbasierte Bestimmung der Modellparameter von Text-Weltmodellen. Dazu gehören sowohl auf den Strukturaufbau textuell kodierter kognitiver Modelle bezogene Parameter der kognitiv orientierten Textlinguistik in quantitativen Operationalisierungen; als auch Parameter der diskursorientierten, parameterbasierten Sprachtypologie und der quantitativen Linguistik (vgl. Kapitel 2), die als textstrukturelle Kenngrößen potentielle TWM-Merkmale darstellen können.

Die Auswahl der Parameter orientiert sich dabei weitgehend an der von Schulze erarbeiteten Systematik von TWM-Parametern (2018; 2019; 2020; vgl. auch 2004a), deren „Komponenten einer Genre-Grammatik“ (Schulze 2020: 607, Tabelle 1) in dieser Arbeit unter Anwendung automatischer Klassifizierungs- und Data-Mining-Methoden (s. Kapitel 4) einer korpusbasierten Feature-Analyse und -Exploration unterzogen werden sollen (Kapitel 6).

3.1 Parameterbestimmung

3.1.1 Textstrukturelle Einheiten und Parametertypen

Eine systematische Bestimmung der verschiedenen Typen von TWM-Parametern erfolgt aus den zuvor in Kapitel 1 getroffenen Annahmen: Ein Text-Weltmodell (TWM) wird hier als ein abstrakt-schematisches kognitives Modell (Weltmodell) bestimmt, das primär strukturelle Eigenschaften eines Typs textueller kognitiver Modelle über quantitative Musterprototypen erfasst, die sich aus den rekurrenten Strukturen der Texte eines Genres im Sinne eines Kommunikationssituationstyps ergeben. Als ein solches kognitives Strukturmodell kann ein TWM mit Schulze (2019: 6, 13; 2018: 18ff.; vgl. 1.1.2) aufgefasst werden als komplexes, schematisch-konstruktionelles sprachliches Zeichen, das auf der Form-Seite aus durch sprachliche Einheiten gebildeten **Textstrukturmustern** besteht (vgl. Heinemann & Viehweger 1991: 170, 174f.; Fix 2008: 66f., 71). Diese repräsentieren auf der Bedeutungsseite des TWM solche für das entsprechende Genre typischen textfunktionalen Eigenschaften, die auf den strukturellen Aufbau der in den Texten dieses Genres kodierten kognitiven Text-Modelle durch Informationseinheiten wie Referenten und Rela-

tionen bezogen sind (s. Schulze 2019: 31).[1] Beispiele für solche TWM-Parameter sind die Topikalitätsstärke (referentielle Strukturmuster, s. Abschnitt 3.6) sowie die Ereignistypik und -abfolge (relationale Strukturmuster, s. Abschnitt 3.7).

Dementsprechend lassen sich für die Operationalisierung dieser quantitativen Text-Modell-Struktur-Parameter zwei grundlegende Typen unterscheiden: **Globale Parameter** sind solche, auf textweite Durchschnitts- und Verhältniswerte der den Text konstituierenden sprachlichen Einheiten bezogene „document-level statistics" (Finn & Kushmerick 2006: 1512) der quantitativen Linguistik, von denen man annehmen kann (vgl. Schulze 2020: 628f.), dass sie als quantitative textstrukturelle Prototypen in der TWM-Schemainduktion durch die Kognition relevant sind: „[…] eine einfache Quantifizierung basaler Text-Parameter wie Anzahl der Tokens, Anzahl der Types, Token-Type-Verhältnis, durchschnittliche Satzlänge usw. […] [*kann*] als Ausgangspunkt für die Herausarbeitung einer Genre-Grammatik dienen" (Schulze 2019: 18).

Diese statistischen Kennzahlen des formalen Textaufbaus beziehen sich also auf die Vorkommenshäufigkeiten und die Frequenzverhältnisse von lexikalischen und grammatischen Einheiten im Text, d. h. auf Frequenzdaten entweder bzgl. des formalen Aufbaus eines Texts durch linguistische Einheiten (also durch Morpheme, Wörter, Phrasen, Clauses, Sätze, Absätze) oder bzgl. Formklassen wie Wortarten oder Phrasentypen. Beispiele solcher globaler Parameter sind etwa die Textlänge als Maß des textuellen Informationsgehalts (vgl. 1.1.3), Type-Token-Verhältnisse als Varianzmaße oder die Lexikalische Dichte als Maß der konzeptuellen Elaboration kognitiver Text-Modelle (s. Abschnitt 3.2). Verschiedene solcher globaler Parameter können als TWM-Teilmodelle in Feature-Sets kombiniert werden (vgl. 1.1.3; s. 4.1.1).

Der zweite TWM-Parametertyp (**textfunktionale Parameter**) bezieht sich auf *explizite* Text-Modell-Parameter, also auf quantitative strukturelle Eigenschaften der – im Text über bestimmte sprachliche Einheiten bzw. allgemeiner über Konstruktionen repräsentierten – Informationseinheiten, aus denen ein kognitives Text-Modell aufgebaut ist (s. Schulze 2019: 20f.; 2020: 606; vgl. auch Schwarz-Friesel & Consten 2014: 58; Schulze 2004a: 551). Gliedern lassen sich diese kognitiven Struktur-Parameter nach den zentralen Strukturbereichen eines solchen Text-Modells, d. h. nach sog. kognitiven Domänen[2] (s. Schulze 2019: 16, 20f.; 2020: 606): Da ein Text-Modell „Referenten als mentale Einheiten mit ihren Relationen und Aktivitä-

1 Vgl. Schulze 2019: 31: „[…] dieses Wissen um Genre-gesteuerte Sprachhandlungen […] ergibt sich unter anderem aus dem Grad der Präsenz und der sequentiellen Verteilung von Einheiten auf der Mikro-Ebene ebenso wie aus den semantischen Spezifikationen der jeweiligen Konstruktionen (im Sinne der Construction Grammar)."

2 Bei Schwarz-Friesel & Consten (2014: 91) werden kognitive Domänen bestimmt als „zusammenhängende Teile von repräsentierten Realitätsbereichen […], d. h. miteinander systematisch verknüpfte Konzepte, die referenzielle Informationen in einem gemeinsamen Netz integrieren."

ten sowie ihrer raumzeitlichen Verankerung speichert“ (Schwarz-Friesel & Consten 2014: 58; vgl. 1.1.2), gehören dazu folglich insbesondere die Bereiche der Raum-Zeit-, Referenten-, Relations- und Informationsstrukturierung.

Diese textfunktionalen, auf die Funktion sprachlicher Einheiten im Strukturaufbau eines Text-Modells bezogenenen TWM-Parameter, die die Anzahl, sequentielle Verteilung, Kategorisierung und Verknüpfung verschiedener Typen von am Aufbau des Modells beteiligten Informationseinheiten erfassen (vgl. Kapitel 1; vgl. auch Tabelle 1 in Schulze 2020: 607), können über entsprechende funktionale linguistische Annotationen berechnet werden: Insbesondere sind dies referenz- und relationssemantische sowie informationsstrukturelle *tags* (s. Abschnitt 5.3). Dementsprechend wird in dieser Arbeit in Analogie zu den Bag-of-Words-Modellen des Information-Retrieval (s. 4.1.1; vgl. auch 2.2.2) bei einfachen, von der sequentiellen Anordnung abstrahierenden Frequenz-Modellen von Informationseinheiten auch von **Bag-of-Tags**-Modellen gesprochen (s. 4.1.1), so etwa bei der Operationalisierung der Ereignistypik eines Textes über die relativen Textfrequenzen der verschiedenen semantischen Verbklassen. Daneben werden aber auch ***tag*-Folgen** untersucht, also sequentielle Verteilungsmuster wie z. B. die semantischer Verbklassen als Operationalisierung typischer Ereignisabfolgen in Text-Modellen.

Weitere Beispiele textfunktionaler Parameter sind Topikalitätsmaße wie die referentielle Distanz, die über Daten zum Auftreten von Referenten im Text berechnet werden, oder informationsstrukturelle Parameter wie Informationsdichte und -fluss, die durch Frequenz- und Abfolgedaten zu Topik-Einführungen berechnet werden können. Auch die textfunktionalen Parameter können ggf. zu Domänen-spezifischen TWM-Teilmodellen oder (auch zusammen mit den globalen Parametern) zu einem TWM-Gesamtmodell kombiniert werden (s. 6.6.1).

Zusammenfassend ergibt sich also folgende Gliederung der Parameter:

- **globale Strukturierung**: Stärke, Verhältnisse und Frequenzverteilungen von Grundeinheiten des sprachlichen Aufbaus (lexikalische, morphologische, syntaktische Einheiten)
- **raum-zeitliche Strukturierung**: Stärke und sequentielle Verteilung (der Kodierung) von Raum- und Zeit-Informationen (Lokalisierungen etc.)
- **Referenten-Strukturierung**: relative Häufigkeit, Frequenzverteilung, durchschnittlicher Abstand und Abfolgestruktur (der Kodierung) von referentiellen Informationseinheiten (Objektvorstellungen, vgl. Schulze 2018: 202)

- **relationale Strukturierung**: relative Häufigkeit, Abfolgestruktur sowie Kookkurrenzmuster (der Kodierung) von (Typen von) relationalen Informationseinheiten (Relatoren und Ereignisvorstellungen)[3]
- **Informationsstrukturierung**: Stärke, regionale Verteilung und Abfolgemuster (der Kodierung) von pragmatischen Informationseinheiten

3.1.2 Quantitative Mustertypen

Wie in Kapitel 1 dargelegt, sind die kognitiven Texttypisierungsparameter primär quantitativer Natur, da typische Strukturen textuell kodierter kognitiver Modelle im TWM-Aufbau als Schemata induktiv als sich wiederholende, also frequenzbezogene Muster des Auftretens von (auch konstruktionellen) sprachlichen Einheiten gelernt werden; vgl. dazu Schulze 2019: 15: „Diejenigen Größen, die in einer Vielzahl von Texten eines vermuteten Genres beobachtet werden, könnten als konstitutiv für dieses Genre betrachtet werden." Diese Bestimmung der TWM-Parameter lässt sich damit grundsätzlich einem *corpus-driven*-Ansatz von Diskursanalysen zuordnen, wie er bei Bubenhofer beschrieben ist (in dieser Arbeit allerdings auf strukturell-schematischer Ebene):[4]

> Typischer Sprachgebrauch kann operationalisiert werden als rekurrentes Auftreten von textuellen Einheiten in bestimmten Sprachausschnitten. Solche textuelle Einheiten können Morpheme sein, Lexeme oder aber komplexe Kombinationen von solchen Einheiten. (Bubenhofer 2008: 408f.)

Dies sind (vgl. auch Kapitel 2) zum einen **paradigmatische Parameter**, die die Häufigkeit des Auftretens von Einheiten im Text (auch im Verhältnis zueinander) in einer oder mehreren Kennzahlen zusammenfassen; dies sind also **Frequenzverteilungen und -verhältnisse** von sprachlichen Grundeinheiten.

3 Relatoren werden hier mit Schulze als das „kognitive Korrelat" (2019: 26) von Verben aufgefasst (s. auch Schulze 2018: 202), Ereignisvorstellungen als das kognitive Korrelat von Clauses als relationalen Basis-Einheiten (s. Schulze 2018: 4). Vgl. Schulze 2020: 619: „Methodisch basieren die hier vorgestellten Annahmen und Beobachtungen auf der These, dass sich einfache Ereignisvorstellungen (EV) sprachlich über ‚einfache Sätze' repräsentieren, die eine primitive Struktur des Typs NP VP [NP] haben."

4 Statt typischer Mehrwortverbindungen wie bei Bubenhofer (2008) werden hier als Textstrukturmuster z. B. typische Verbklassenfolgen aus dem Korpus extrahiert.

Zum anderen sind hier verschiedene Arten von **syntagmatischen Verteilungen** relevant, die sich auf die lineare Abfolgestruktur von textstrukturellen Einheiten beziehen;[5] zu diesem sequentiellen Parametertyp gehören:

- **regional differenzierte Frequenzverteilungen** (regionaler Kontext)
- **Kookkurrenzen** im Sinne von Mustern häufiger Abfolgen (Auftreten im selben Kontext)
- **Partitur-Folgen** als Sequenzen von Frequenzwerten von Einheiten bzgl. ihres Auftretens in einer Basis-Textstruktur-Einheit (z. B. Clauses, s. u.)

Abbildung 3.1: Beispiel einer Partitur-Folge

Der von Weinrich (1976: 145ff.) eingeführte Begriff der **Textpartitur** bezeichnet dort ein „Notationsverfahren der Textanalyse, das in Form einer partiturähnlichen Matrix jedem Textsegment seine grammatischen Merkmale zuordnet, sodass sich an der Wiederholung oder Veränderung von Merkmalen (z. B. der Übergang beim Wechsel von Aktiv zu Passiv) Struktureigenschaften des Textes ablesen lassen" (Bußmann 2008: 725f.). Eine solche Partitur-Folge (vgl. Abbildung 3.1) legt gewissermaßen den Rhythmus des Auftretens einer Einheit im Text bzgl. der „Abfolge der grundlegenden Einheiten eines Textes" (Schulze 2019: 13) fest – insbesondere bzgl. der Abfolge von Clauses als linguistischer Repräsentation von Ereignisvorstellungen als textuelle Basiseinheit (Schulze 2019: 10, 13). Solche Partitur-Operationalisierungen verbinden also eine syntagmatische mit einer paradigmatischen Quantifizierung, indem sie den Verlauf der Stärke einer Einheit in den durch die Basiseinheit festgelegten Schritten der linearen Textabfolge abbilden. Bei Schulze bilden solche Partitur-Folgen, die in ihrer Gesamtheit als „Partiturstimmen" (Schulze 2020: 629f.) die Gesamt-Partitur eines Textes ergeben, das zentrale Repräsentationsmittel für sequentielle TWM-Muster-Parameter:

5 S. Legallois & Charnois & Larjavaara (2018a: 2) zur Unterscheidung von paradigmatischen Parametern der klassischen quantitativen Textlinguistik von syntagmatischen Parametern in der Genre-Analyse; vgl. dazu auch Schulze 2019: 6 (Hervorhebung im Original): „Demnach ist von einem jeweils für ein Genre einschlägigen Satz an ‚Ausdruckstypen' auszugehen (etwa Aktanz, Zeit, Raum, lexikalische Komplexität usw.), die mit entsprechenden konzeptuellen Einheiten gekoppelt sind […]. Da Humana nur ‚in der Zeit' […] *handeln* können, werden auch die Typen einer Genre-gesteuerten Handlung zumindest minimal linearisiert, wodurch sich weitergehende Typen ergeben, die mit der Semiotik der linearen Charakteristik von Handlungen bzw. Handlungssequenzen verbunden sind (wie Kausalität, referentielle Kontiguität usw.). Das Gesamt dieses strukturellen Geflechts von Types kann als *Textur eines Genres* bezeichnet werden, die die Topologie (oder ‚Syntax') eines Genres abbildet."

> Der Begriff Partitur soll einerseits dem Fakt Rechnung tragen, dass Texte analog zu Partituren sequentiell aufgebaut sind, aber zugleich nicht-lineare Komponenten beinhalten. Zugleich ergibt das Gesamt einer Partitur in ihrer Verknüpfung von substantiellen und strukturellen (schematischen) Einheiten eine einheitliche Gestalt, deren Bedeutung das TWM eines Textes ist. […] Natürlich handelt es sich bei der Partitur nicht um ein direkt substantiell repräsentiertes Zeichen, sondern um das, was in der Construction Grammar als *fully schematic construction* bezeichnet wird. Es geht also um ein rein schematisches oder strukturelles Zeichen, das sich über die Typizität und Sequenz der in einem Text gegebenen sprachlichen Zeichen äußert. (Schulze 2019: 13)

Wie oben erwähnt, werden in dieser Arbeit neben solchen Partitur-Folgen aber auch andere Repräsentationstypen sequentieller Strukturinformationen erprobt, etwa textweite kategoriale *tag*-Folgen, häufige Abfolgemuster solcher *tag*-Folgen (Frequent-Patterns), die Operationalisierung regionaler Verteilungsmuster über *bag*-Features oder auch die Berechnung textweiter sequentieller Merkmale für Informationseinheiten wie z. B. der Abstand von Referenten über den Parameter der referentiellen Distanz.

Die verschiedenen Typen der in dieser Arbeit verwendeten quantitativen TWM-Parameter werden in Tabelle 3.1 anhand des Merkmals der Clauselänge als mögliches Maß der kognitiven Elaboration von Ereignisvorstellungen exemplifiziert; wie oben angemerkt, bietet sich für einen Parameter meist ein bestimmter Operationalisierungstyp an. Datengrundlage des Beispiels ist der obugrische Text 741 des Korpus (s. Interlinearversion 2 in 5.3.6), der lediglich aus vier Clauses besteht. Im Beispiel wird die Clauselänge über die Tokenanzahl pro Clause operationalisiert (vgl. 3.2.3; Clauselängen im Beispiel: 5, 1, 4, 1):

Durchschnittliche Clauselänge pro Text:							
$\frac{\text{Tokenanzahl des Textes}}{\text{Clauseanzahl des Textes}} = \frac{5+1+4+1}{4} = \frac{11}{4} = 2.75$							
Frequenzverteilung von Clauselängen:							
Clauselänge (Tokenanzahl):	1	2	3	4	5	6	…
Häufigkeit von Clauses mit entsprechender Länge:	2	0	0	1	1	0	…
Regionale Verteilung von Clauselängen:							
Textregionen:	Region 1:		Region 2:				
durchschnittliche Clauselänge pro Region:	$\frac{5+1}{2} = 3.0$		$\frac{4+1}{2} = 2.5$				
Frequent-Patterns als Kookkurrenzen von Clauses bestimmter Längen:							
Teilsequenzen von Clauselängen:	`5-1`	`1-4`	`4-1`	`5-1-4`		`1-4-1`	
relative Häufigkeit des Musters:	0.33	0.33	0.33	0.5		0.5	
Partitur-Folge der Clauselängen:							
Sequenz der Tokenanzahl pro Clause: `5-1-4-1`							

Tabelle 3.1: Operationalisierungstypen am Beispiel Clauselänge

3.1.3 Systematik der Parameter

Abschließend erfolgt hier in Tabelle 3.2 eine systematische Auflistung über die in den jeweiligen Parametertypbereichen (kognitiven Domänen) als Kodierungsmittel entsprechender Informationseinheiten und Strukturmuster relevanten linguistischen Einheiten und Ausdrucksmittel. Dazu gehören insbesondere: syntaktische, semantische und pragmatische Rollen; semantische Klassen; phorische Relationen (Referentialisierung); Sprechakttypen; Lokalisierung, Temporalisierung, Modalisierung (vgl. Schulze 2000b: 122). Die konkret für die obugrischen Sprachen für die Korpusauswertung in Kapitel 6 relevanten Ausdrucksmittel werden in Abschnitt 5.1.2 besprochen. Tabelle 3.2 gibt auch eine Übersicht darüber, welche der besprochenen Quantifizierungsmethoden sich für die Operationalisierung der unterschiedlichen TWM-Parameterbereiche eignen.

Kognitive Domäne	Quantitative Modelle	Linguistische Einheiten	Quantitative Merkmale
Globale Strukturierung	Genre-Typik allgemein-formaler Strukturierung	Lexikalische, grammatikalische und syntaktische Einheiten	Durchschnittswerte, Frequenzverteilungsmuster
Raum-zeitliche Strukturierung	Genre-Typik adverbialer Funktion (Lokalisierung, Temporalisierung)	Adverbiale Markierung (Adpositionen, Lokalkasus), TAM-Markierung (Tempus, Aspekt, Modus)	Durchschnittswerte, Frequenzverteilungsmuster, sequentielle Verteilungsmuster (Partitur-Folgen)
Referenten-Strukturierung	Genre-Typik referentieller Funktion (Kodierung von Objektvorstellungen)	Nominale, pronominale und phorische Einheiten	Durchschnittswerte, Frequenzverteilungen
Relationale Strukturierung	Genre-Typik relationaler Funktion (Kodierung von Relatoren bzw. Ereignisvorstellungen)	Verben, semantische Verbklassen	Frequenzverteilungsmuster, sequentielle Verteilungsmuster (Partitur-Folgen), Kookkurrenz-Muster (Frequent-Patterns)
Informationsstrukturierung	Genre-Typik pragmatischer Funktion (Kodierung schematisch-textstruktureller Funktion)	Topik- und Fokusmarker, Wortstellung, Anaphorik, Diathesen, Definitheitsmarker, Switch/Same-Subjekt-Marker und -Konstruktionen	Regionale Verteilung, sequentielle Verteilungsmuster (Partitur-Folgen)

Tabelle 3.2: Übersicht zur Systematik der Parameter

Im Folgenden wird eine knappe Übersicht über die in dieser Arbeit untersuchten quantitativen kognitiven TWM-Parameter und ihre konkrete Operationalisierung gegeben, bevor diese Parameter im weiteren Verlauf dieses Kapitels (ab Abschnitt 3.2) dann im Einzelnen vorgestellt und theoretisch verortet werden.

Ausgewählt wurden dabei solche quantitativen textstrukturellen Parameter der kognitiven Textlinguistik, der quantitativen Linguistik sowie der diskursanalytisch orientierten Sprachtypologie, die eine Relevanz bzgl. der Genre-Differenzierung erwarten lassen. Wie oben festgestellt, orientiert sich diese Auswahl hier vor allem an der Systematik der Parameter zu einer TWM-Operationalisierung von Schulze (s. insbesondere 2020: 607). Für die **globalen**, textstatistischen TWM-Parameter mit ihrer kognitiv-textlinguistischen Interpretation (Inferenz, Elaboration usw.) werden hier primär die für die Analyse einzelner udischer Volkserzählungen erarbeiteten Parameter in Schulze 2018 und 2004a herangezogen. Für die **textfunktional** motivierten Parametertypen sind vor allem die Operationalisierungen in Schulze 2019 bzw. 2020 maßgebend – allerdings in einer für die automatische, korpusbasierte Auswertung geeigneten Anpassung (vgl. Abschnitt 1.2; z. B. werden die Hauptreferenten eines Textes automatisch durch Frequenzsortierung berechnet, s. 3.6.1).

Globale Strukturierung:

- **Varianz bzw. Redundanz**: Type-Token-Verhältnisse
- **Lexikalische Dichte**: Anteil von Inhaltswörtern
- **Elaboration**: Länge sententieller oder phrasaler Einheiten
- **Komplexität**: Stärke eingebetteter Einheiten
- **Explizitheit**: Stärke lexikalischer Einheiten
- **Inferenz**: Anteil nicht-overt kodierter Einheiten

Raum-zeitliche Strukturierung:
Spatiotemporale TWM-Eigenschaften werden nicht als gesonderte Parameter operationalisiert, sondern insbesondere über folgende Parameter mitbehandelt (s. auch Abschnitt 3.5):

- **Referentielle Distanz und Topik-Persistenz**
- **Ereignisabfolge**
- **Temporal-Sequencing**

Referenten-Strukturierung:

- **Referentielle Distanz**: Abstand koreferentieller Einheiten
- **Topik-Persistenz**: Anzahl verbleibender Erwähnungen einer referentiellen Einheit
- **Topikalitätsquotient**: Frequenzverteilung von Referenten

Relationale Strukturierung:

- **Ereignistypik**: Stärke von Relationstypen
- **Ereignisabfolge**: Verteilung Relationstypen
- **häufige/typische Ereignismuster**: Kookkurrenz von Relationstypen

Informationsstrukturierung:

- **Informationsdichte und -fluss**: Stärke und Verlauf der Einführung neuer Topiks
- **Perspektivierung** (**Switch-Reference**): Verteilung von Referentenerwähnungen über den Text (Thematische Progression)
- **Aufmerksamkeitssteuerung**: Stärke und Typik von Fokussierungsstrategien
- **Vordergrund-Hintergrund-Struktur**: Verteilung von subordinierten Einheiten als Hintergrundinformation (**Komplexität**), Verteilung von Präsensformen als Vordergrund-Marker (**Temporal-Sequencing**)
- **Textinterne Diskursstrukturen**: Stärke und sequentielle Verteilung direkter Rede

3.2 Globale Genre-Parameter

Die globalen textstrukturellen Parameter für eine TWM-Operationalisierung sind textweite quantitative Durchschnitts- und Verhältnisdaten der den Text konstituierenden sprachlichen Einheiten. Zu diesen **textstatistischen Genre-Parametern** gehören insbesondere Stilistik-Maße der quantitativen Linguistik, etwa verschiedene Varianten von Type-Token-Verhältnissen als Maße für lexikalische Varianz bzw. Redundanz, die also Rückschlüsse auf die konzeptuelle Variabilität geben können (vgl. Schulze 2019: 187) und damit potentiell Genre-differenzierende Merkmale darstellen können (vgl. Schulze 2020: 610). Daneben werden weitere klassische textstatistische Merkmale, wie sie als Parameter in der quantitativen Textlinguistik Anwendung finden (s. Mehler 2005: 333), als potentielle TWM-Parameter berücksichtigt; dazu gehören insbesondere Satz- und Clauselängenmaße, etwa die Phrasenanzahl pro Clause als Maß der konzeptuellen Elaboration von Ereignisvorstellungen.

Die in den Formeln der folgenden Abschnitte verwendeten Kürzel (z. B. `RED` für Redundanz) beziehen sich auf die in der Auswertung in Kapitel 6 verwendeten Abkürzungen für die Feature-Variablen (vgl. die Feature-Abkürzungen im Abkürzungsverzeichnis).

3.2.1 Varianz und Redundanz

Allgemein geben **Type-Token**-Verhältnisse als die mittlere Frequenz der verschiedenen Vorkommen (Token) einer bestimmten Einheit (Type) Auskunft über die **Varianz** bzw. (als inverses Verhältnis) die **Redundanz** des Vorkommens einer Einheit in einem Text, also darüber, wie oft sich eine Einheit im Text im Durchschnitt wiederholt:

$$\text{Varianz} = \frac{\text{Typeanzahl}}{\text{Tokenanzahl}} \tag{3.1}$$

$$\text{Redundanz (RED)} = \frac{\text{Tokenanzahl}}{\text{Typeanzahl}} \tag{3.2}$$

In einer einfachen, wortformbezogenen Type-Definition (jede Wortform ist ein Type) ist dies in erster Linie ein sprachtypologisches Maß von morphologischer Varianz, das abhängig ist von der morphosyntaktischen Typologie der relevanten Sprache (s. Schulze 2018: 186; 2020: 612).

In einer lemmabezogenen Type-Definition, in der die Menge aller Wortformen eines Lexems einen Type bildet (vgl. Schulze 2020: 612), ist dies in der quantitativen Linguistik ein Maß von **lexikalischer Varianz** bzw. Redundanz (vgl. Kapitel 2). Eine durch das wiederholte Vorkommen derselben Informationseinheiten auftretende hohe lexikalische Redundanz ist nach Schulze der anzunehmende Prototyp für Volkserzählungen, was kognitiv zu interpretieren ist als „Aufbau eines geschlossenen referentiellen Wissens" (Schulze 2020: 607).

In Kapitel 6 wird zunächst das einfache, wortformbezogene Token-Type-Verhältnis als allgemeiner Redundanz-Parameter getestet; da hier Texte eng verwandter Sprachen (also mit sehr ähnlichem morphologischem Sprachbautyp) untersucht werden, macht es durchaus Sinn, dieses Maß als potentiell Genre-differenzierendes Merkmal mitzuberücksichtigen. Die lemmabezogene Variante wird als lexikalische Dichte im folgenden Abschnitt behandelt.

3.2.2 Lexikalische Dichte

Maße der **lexikalischen Dichte** (auch: *lexical diversity* oder *lexical richness*, s. Daller 2010) beziehen sich gewöhnlich auf den Anteil von Inhaltswörtern eines Textes; sie bestimmen also das Verhältnis von Inhaltswörtern zu Tokenanzahl.

$$\text{Lexikalische Dichte (allgemein)} = \frac{\text{Lexemanzahl (Inhaltswörter)}}{\text{Tokenanzahl}} \tag{3.3}$$

Für eine kognitive Interpretation im Zusammenhang mit konzeptueller Varianz kann das Type-Token-Verhältnis von Inhaltswörtern berechnet werden, normiert bzgl. der Textlänge: „Lexical bases gives [*sic*] us the starting point to calculate the conceptual variability of a text" (Schulze 2018: 187). Die lexikalische Varianz bzw. Dichte kann auch als Maß konzeptueller Elaboration aufgefasst werden, vgl. dazu Schulze 2019: 19: „Die Zahl der lexikalischen Basen (LB) sowie das entsprechende Token/LB-Verhältnis verdeutlichen den Grad der konzeptuellen Elaboration. Vereinfacht gesagt: Im Gegensatz zu den Evangelien bedienen sich die Erzählungen eines relativ kleinen lexikalischen Inventars."

Zur Normierung der Werte bzgl. unterschiedlicher Textlängen für einen Vergleich der lexikalischen Varianz von Texten wurden in der quantitativen Linguistik verschiedene Maße aufgestellt (so z. B. das MTLD-Maß als *measure of textual lexical*

diversity, s. Mehler 2005: 340, 344; oder eine Normierung durch den Guiraud-Index, s. Daller 2010; vgl. Schulze 2019: 19; 2018: 187). In Kapitel 6 wird eine Variante mit lemmabezogener Type-Definition sowie Wurzel-Textlängennormierung durch den Guiraud-Index verwendet:

$$\text{Lexikalische Dichte (LEX\_DENS)} = \frac{\text{Lemma-Typeanzahl}}{\sqrt{\text{Tokenanzahl}}} \tag{3.4}$$

3.2.3 Clause-Elaboration

Ein Maß, das in der quantitativen Stilistik für die Differenzierung von Textsorten häufig Anwendung findet (Mehler 2005: 333), ist die **Clauselänge**, in der einfachsten Version als durchschnittliche Tokenanzahl pro einfachem Satz. Allgemein kann die durchschnittliche Länge von sprachlichen Einheiten, also die Anzahl der diese konstituierenden Einheiten, als Maß ihrer Elaboration dienen (vgl. Schulze 2018: 188). Entsprechend kann man auch die Clauselänge interpretieren, vgl. Schulze 2019: 20: „Die Ratio Tokens/ES [*ES = einfacher Satz*] informiert darüber, in welchem Umfang die einzelnen Ereignisvorstellungen elaboriert werden."

$$\text{Clause-Elaboration (allgemein)} = \frac{\text{Tokenanzahl}}{\text{Clauseanzahl}} \tag{3.5}$$

Auch dieses einfache Maß ist allerdings (ähnlich wie die Type-Token-Verhältnisse) abhängig vom morphologischen Sprachbautyp (vgl. Schulze 2018: 188). Für die Berechnung der lexikalischen Clause-Elaboration als kognitiven TWM-Parameter schlägt Schulze daher – analog zur Operationalisierung der lexikalischen Dichte – ein über den Guiraud-Index textlängennormiertes Maß der durchschnittlichen Anzahl lexikalischer Basen pro Clause vor (Schulze 2018: 188f.), das entsprechend als Maß der **konzeptuellen Elaboration** der im Clause versprachlichten Ereignisvorstellung interpretiert werden kann (s. Schulze 2019: 19).

Alternativ wird in dieser Arbeit als längenbasiertes Maß der konzeptuellen Elaboration von durch Clauses kodierten Ereignisvorstellungen die durchschnittliche Phrasenanzahl pro Clause berechnet, also aus wie vielen Satzgliedern (und damit Informationseinheiten) ein Clause besteht:

$$\text{Konzeptuelle Clause-Elaboration (CL\_ELAB)} = \frac{\text{Phrasenanzahl}}{\text{Clauseanzahl}} \tag{3.6}$$

3.2.4 Komplexität

Komplexität wird hier allgemein aufgefasst als die Stärke eingebetteter Einheiten; als weiteres globales textstrukturelles Maß kann also der Anteil subordinierter Einheiten pro Satz (insbesondere subordinierter Clauses) bzw. auch pro Clause (z. B. von

Infinitiv-Komplementen) als textweites Maß für *Komplexität* gelten (Givón 1983a: 23f.; Schulze 2004a: 555):

$$\text{SENT\_COMPLEX} = \frac{\text{Anzahl subordinierter Clauses/Verbalkonstruktionen}^{6}}{\text{Satzanzahl}} \tag{3.7}$$

$$\text{CL\_COMPLEX} = \frac{\text{Anzahl subordinierter Verbalkonstruktionen}}{\text{Clauseanzahl}} \tag{3.8}$$

Die diskurstypologische Forschung sieht einen engen Zusammenhang von Subordination mit **Backgrounding**, also der Kodierung von Hintergrundinformationen, vgl. etwa Givón 1983a: 23: „[…] main clauses carry the bulk of sequentially-ordered new information in discourse, and various subordinate clauses may carry discontinuous, non-sequential background information." Die Komplexität im Sinne syntaktischer Einbettung kann dementsprechend als informationsstruktureller Parameter der Backgrounding-Stärke interpretiert werden (s. Schulze 2018: 214f.; Schulze 2004a: 556).[7]

Die syntaktische Komplexität gibt also Hinweise auf thematisch-pragmatische Komplexität in der Versprachlichung, d. h. auf die informationsstrukturelle Bedeutungsstruktur des Textes; entsprechend wird in Kapitel 6 neben der durchschnittlichen Stärke subordinierter Einheiten auch der Komplexitätsverlauf im Rahmen der auf informationsstrukturellen Annotationen basierenden Operationalisierungen als Sequenz-Parameter modelliert, also als *foreground-background*-Abfolge-Struktur.

3.3 Globale referenzbezogene Genre-Parameter

Als globale Parameter für eine referentenbezogene kognitive Texttypologie werden Parameter diskutiert, die sich auf quantitative Aspekte der Benennung und Versprachlichung von Referenten beziehen: Wie stark werden Referenten vom Sprecher explizit benannt? Welche referentenbezogenen Informationen werden vom Hörer durch Weltwissen ergänzt, um Kohärenz herzustellen? Diese Parameter beziehen sich also allgemein auf die konzeptuelle (d. h. auf Informationseinheiten bezogene)

6 Die Berechnung sowohl der Satz- als auch der Clause-Komplexität geschieht in Kapitel 6 über die Daten zu den im Korpus als subordiniert ausgezeichneten verbalen Einheiten (SUBPRED; s. 5.3.2). Die Differenz zwischen Satz- und Clause-Komplexität ist dann nur die Bezugsgröße im Nenner.

7 Im Zusammenhang mit der kognitiven Sprachverarbeitung syntaktischer Einbettung ist auch die psycholinguistische Studie von Karlsson 2007 zu beachten, in der dieser die maximale Rekursionstiefe von *center-embedding*-Strukturen in Korpora untersucht und zu dem Ergebnis kommt, dass diese aufgrund von Begrenzungen der Kapazität des Arbeitsgedächtnisses von der menschlichen Kognition nur zu einer begrenzten Tiefe verarbeitet werden können.

Elaboration von Referenten im kognitionspsychologischen Sinn als Hinzufügung von Informationen:

> In order to elaborate the linguistically encoded meaning of a text (which is based on the grammatical and lexical text structure), recipients automatically construct a mental text-world model [*hier im Sinne des konkreten, durch einen Text kodierten kognitiven Text-Modells*]; i. e. recipients enrich the text base by incorporating both information from the text and information activated through conceptual instantiation and inferential processing. (Schwarz-Friesel & Consten 2011: 351)

Einerseits ist hier also bzgl. der Elaboration von referentiellen Repräsentationen die Frage relevant, wie stark der Text selbst im referentiellen Bereich elaboriert ist, also nach der Stärke der **qualifizierenden Elaboration** von Einheiten, dem Grad der **Explizitheit** der Benennung usw., andererseits die Frage, wie stark das textuell kodierte kognitive Modell noch vom Rezipienten um Referenten-Einheiten elaboriert werden muss, die nicht textuell kodiert sind.

Diese zweite Elaborationsart betrifft also Elaboration im Sinne einer kognitiven Erweiterung der textuell kodierten Informationen im Textverarbeitungsprozess aus verschiedenen, im Langzeitgedächtnis des Rezipienten abgespeicherten Wissensquellen (s. Schwarz-Friesel & Consten 2011: 352; vgl. auch Ungerer & Schmid 1996: 160ff.). Dabei kann man diesen Typ der Elaboration – also die Ergänzung von Informationseinheiten durch Schließen aus Weltwissen (s. Ungerer & Schmid 1996: 213ff. zur Ergänzung von Ereignisabfolgen; vgl. auch Schwarz-Friesel & Consten 2014: 70; DeLamater & Myers & Collett 2018: 212) – als **Inferenz** von der **Unterspezifikation** als der ersten Elaborationsart unterscheiden (Schwarz-Friesel & Consten 2011: 350f., 361). Mit Unterspezifikation ist die Ergänzung typischer Referenten gemeint, die basierend auf Sprachwissen bzgl. semantischer Rollen und entsprechend verknüpftem Wissen über Default-Werte dieser Rollen durchgeführt wird (z. B. *Messer* als Default-Wert der Instrument-Rolle von *schneiden*, s. Schwarz-Friesel & Consten 2014: 66f., 71; vgl. auch Ungerer & Schmid 1996: 160ff.).

Man kann davon ausgehen, dass die quantitativen Ausprägungen der Elaboration referentieller Einheiten von Text-Modellen (etwa der Anteil versprachlichter vs. erst im Modellaufbau hinzugefügter Referenten des Text-Modells oder der Grad der expliziten Qualifikation von Referenten) mit der Art der Kommunikationssituation variieren und diese also potentielle Parameter von TWM-Genre-Modellen darstellen (vgl. Schulze 2020: 614ff.; 2019: 26f.; 2018: 201ff.; 2004a: 561f.). Insbesondere erlaubt die Berechnung der durchschnittlichen Frequenz von Inferenz-bezogenen Elaborationen für die Texte eines Genres die Abschätzung des in der Konstruktion von Text-Modellen dieses Texttyps erlaubten bzw. erwartbaren Grads des Einbeziehens von Weltwissen (also von anderen **Weltmodellen**, d. h. episodischen, allgemein-enzyklopädischen oder kulturellen Wissenseinheiten; vgl. Dik 1997b: 411f.).

Für die Operationalisierung dieser referenzbezogenen globalen Parameter sind insbesondere die Realisierungsarten des sprachlichen Ausdrucks von Referenten (also von nominalen Einheiten) zentral:

Referentieller Ausdruck	Kodierung substantiell	Kodierung overt	Phrasentyp
lexikalisch	+	+	NP
pronominal	+	partiell	PRONP
Nullanapher	–	–	Ø

Tabelle 3.3: Realisierungstypen referentieller Ausdrücke

Für die Berechnung von referentiellen Elaborationstypen als quantitative TWM-Modell-Parameter ist also eine Identifizierung von nicht-overt kodierten Referenten, d. h. von Nullanaphern, notwendig (vgl. auch die Berechnung referentieller Dichte über Argument-Positionen in 3.3.2). Im obugrischen Korpus liegt eine direkte Auszeichnung von Nullanaphern für die zentralen syntaktischen Funktionen Subjekt und Objekt vor, die in einer halbautomatischen Annotation erstellt wurde (s. Abschnitt 5.3). Alternativ ist auch die Verwendung von Informationen zu Subkategorisierungsrahmen wie aus der FrameNet-Datenbank denkbar (vgl. Chambers & Jurafsky 2009), um automatisch Rollen-bezogene Auslassungen zu extrahieren.

3.3.1 Referentieller Inferenzgrad

Der **referentielle Inferenzgrad** bezieht sich auf den Anteil nicht-overt kodierter referentieller Einheiten im Text (vgl. Schwarz-Friesel & Consten 2014: 70; Schulze 2004a: 556, 561f.). Dieser lässt sich entsprechend über den Anteil von Nullanapher-kodierten Referenten berechnen:

$$\text{REF\_INFER} = \frac{\text{Anzahl Nullanaphern (nicht-overt)}^{8}}{\text{Anzahl referentieller Einheiten}} \tag{3.9}$$

Von dieser Operationalisierung inferierter Einheiten über Nullanaphern ist anzunehmen, dass sie den allgemeinen Grad von **Wissensergänzung** referentieller Informationseinheiten abbildet (also, wie viele Referenten im Durchschnitt nicht versprachlicht sind und dementsprechend von der Kognition elaboriert werden müssen), ohne dass damit eine explizite Trennung von Inferenz im engeren Sinne (Erschließen von Referenten aus Weltwissen durch Inferenzschlüsse) und Unterspezifikation (als Rollenergänzung aus Sprachwissen) verbunden ist.

Allerdings ist die Verwendung von Nullanaphern sprachspezifisch unterschiedlich grammatikalisch determiniert. So könnte diese Operationalisierung über den Grad

8 In Kapitel 6 erfolgt deren Bestimmung über getaggte Nullmorpheme (`zero`) und referentiell spezifizierende Possessivsuffixe (`px`; s. dazu 5.1.2 sowie die Interlinearversion 1 in 5.3.6 für ein Beispiel).

der Verwendung von Nullanaphern – etwa bei pro-drop-Sprachen wie den obugrischen Sprachen, bei denen die Realisierung von topikalen Referenten in Subjekt- und auch in Objekt-Position nicht obligatorisch ist (die Verwendung der Nullanapher ist hier also vor allem grammatikalisch-pragmatisch determiniert) – eher den Grad der Referenz auf bekannte, topikale Referenten im Text anzeigen (also ein informationsstruktureller Parameter sein; vgl. Abschnitt 3.8 zu Topik-Einführung und Perspektivierung). Zumindest ist davon auszugehen, dass dadurch die Anzeige von inferentieller Elaboration über Nullanaphern verzerrt ist. Inwiefern der Parameter dennoch sinnvoll zur Genredifferenzierung eingesetzt werden kann, soll in Kapitel 6 anhand des Erprobungskorpus untersucht werden.

Zusätzlich zu diesem Maß referentieller Inferenz über den Anteil von Nullanaphern als nicht-overt kodierten referentiellen Einheiten wird in Kapitel 6 für die Texte des obugrischen Korpus auch der Anteil von partiell-coverten referentiellen Einheiten berechnet, also von pronominalen Anaphern. Je nach einzelsprachlichem Anaphorik-System – etwa bei syntaktisch geforderten pronominalen (expletiven) Ergänzungen – kann diese Metrik den **pronominalen Inferenzgrad** angeben bzw. Auskunft über die Explizitheit der Benennung geben (vgl. 3.3.4):

$$\text{PRON\_INFER} = \frac{\text{Anzahl pronominaler referentieller Einheiten (partiell-covert)}}{\text{Anzahl referentieller Einheiten}} \tag{3.10}$$

3.3.2 Referentielle Dichte

In dieser Arbeit werden unterschiedliche, in der sprachtypologisch-diskursfunktionalen Literatur gegebene Operationalisierungen **referentieller Dichte** (Du Bois 1987: *information pressure*) bzgl. ihrer Eignung als TWM-Genre-Parameter überprüft; an dieser Stelle – im Rahmen der textstatistisch über syntaktisch-morphologische Eigenschaften berechenbaren, globalen referentiellen Parameter – zunächst im Sinne von Bickel (2003: 726) über den Anteil overter NP-Argumentrealisierungen bzgl. Argument-Positionen.[9]

$$\text{REF\_DENS} = \frac{\text{Anteil overter (lexikalischer) Argument-NPs}}{\text{Anzahl zentraler Argument-Positionen}} \tag{3.11}$$

In 3.8.1 wird dann die relative referentielle **Informationsdichte** als informationsstruktureller Parameter über Eigenschaften der textuell kodierten Informationseinheiten operationalisiert, nämlich über den Anteil neuer Topiks (Biber 1992b; Du Bois 1987; 2003). Während sich die Berechnung der Informationsdichte dort auf alle Referen-

9 In Kapitel 6 erfolgt die Berechnung auch inkl. partiell overten nominalen Einheiten, also Nomen und Pronomen (vgl. Bickel 2003: 721). Im Obugrischen als pro-drop-Sprache ist die Verwendung pronominaler Einheiten stark eingeschränkt; wenn nicht pragmatisch markiert, werden bekannte Referenten über Nullanaphern kodiert (s. Kapitel 5).

tenerwähnungen als Grundgesamtheit bezieht, beschränkt sich die morphologisch-syntaktische Operationalisierung der referentiellen Dichte über den Anteil lexikalischer NPs auf die Realisierungstypik der zentralen, verbgeforderten Argument-Positionen (vgl. Bickel 2003: 721), d. h. auf die Argumentstruktur als „architecture *for* cognitive processing, in which certain locales are predictably specialized for high- or low-cost work" (Du Bois 2003: 81, Hervorhebung im Original). Dies gilt insbesondere für die Subjekt- und Objekt-Positionen in transitiven Sätzen, für deren Realisierung Du Bois 1987 mit der **Preferred Argument Structure** informationsstrukturelle Constraints aufstellt (vgl. Cumming & Ono & Laury 2011), nämlich dass nicht mehr als ein neuer Referent pro Clause in den beiden Argument-Positionen auftritt (Du Bois 1987: 819) und dass neue Referenten nicht in A-Position auftreten („Non-new A Constraint", Du Bois 1987: 819), d. h. in der Position des agensartigen Arguments eines transitiven Satzes (in Akkusativsprachen dessen Subjekt).

Diese sprachunabhängigen, pragmatisch begründeten (also im Sinne dieser Arbeit aus dem Aufbau kognitiver Text-Modelle folgenden) Beschränkungen für die Anzahl und die Art von Referenten in zentralen Argument-Positionen in transitiven Clauses geben gleichzeitig auch Grenzen für deren referentielle Dichte vor:

> [...] information distribution among argument positions in clauses of spoken discourse is not random, but grammatically skewed toward an ergative pattern. Arguments comprising new information appear preferentially in the S or O roles, but not in the A role – which leads to formulation of a Given A Constraint. (Du Bois 1987: 805)

Neben dem Informationsstatus der Referenten (neue vs. bekannte Information) ist nach Du Bois auch deren **Topikalitätsstatus** ein informationsstrukturelles Kriterium für Präferenzen in der Kodierung syntaktischer Relationen (s. 3.6.1 zum Parameter des Topikalitätsquotienten). Als konkurrierende pragmatische Motivation für die Kodierung der zentralen syntaktischen Funktionen führt Du Bois (1987: 843) die **Topic-Continuity** an, die durch Grammatikalisierung zu akkusativischen Strukturen führt. Ebenso wie das Non-new-A-Constraint bezieht sich auch die Topikkontinuität hier primär auf eine informationsstrukturelle Eigenschaft bzw. Tendenz von kognitiven Text-Modellen, nämlich auf die Existenz eines über mehrere Sätze kontinuierlichen Topiks (gleichbleibender Referent als Ausgangspunkt einer Kette von Handlungen) im Sinne eines Diskurstopiks als zentralem Thema des Textes bzw. Textteils, also eines Referenten mit hohem „agency potential" (Du Bois 1987: 844), der in transitiven Sätzen entsprechend die Rolle des Subjekts als dem agensartigen Argument einnimmt (*same subject*, vgl. 3.8.2; vgl. auch Li & Thompson 1976: 484: „Subjects are essentially grammaticalized topics").

Diese informationsstrukturellen Constraints des referentiellen Aufbaus von Text-Modellen sind grammatikalischen Strategien zu ihrer Kodierung vorgelagert (vgl. Du Bois 1987: 843, 850) – etwa der Verwendung von Diathesen, um die topikalen Haupt-

referenten syntaktisch in Subjektposition zu halten, falls ihre semantische Rolle in einem Clause nicht der agentivischen Default-Rolle als Diskurstopik entspricht.

Man kann davon ausgehen, dass die beiden informationsstrukturellen Strukturmerkmale der Topikkontinuität sowie der referentiellen Dichte in der Stärke ihrer Ausprägung mit dem Texttyp variieren, sie also als Genre-Parameter herangezogen werden können:

> Information pressure apparently correlates with discourse genre. In some genres, pressure is often high – such as 3rd person stories about strangers, as in the Pear Film narratives. In others, information pressure is often low – such as intimate conversation between family members or long friends, where interlocutors may refer to each other with 1st and 2nd person pronouns, and otherwise share large amounts of currently active background information. (Du Bois 1987: 835)

Der in einem Text-Modell aufscheinende Grad von Topikkontinuität, ebenso wie die Stärke der referentiellen Informationsdichte, hängt also von der für ein Genre typischen referentiellen kognitiven Strukturierung ab: In narrativen Texten ist von einem handlungsbezogenen kognitiven Text-Modell mit stark topikalen, die Handlung tragenden (agentivischen) Referenten in A-Position als kontinuierlichen Topiks auszugehen (d. h. von einer hohen Topikkontinuität) sowie im Vergleich mit informativen Texten wie Zeitungstexten von einer eher niedrigen Informationsdichte. Dies gilt insbesondere für längere Erzählungen (vgl. Du Bois 1987: 835), in denen es mehrere Diskurstopiks gibt, die miteinander (in transitiven Ereignisvorstellungen) interagieren (also auch regelmäßig die eigentlich für neue Referenten präferierte Objekt-Position einnehmen).[10] Für informative Texte ist dagegen von einer hohen referentiellen Informationsdichte, also vielen neuen Referenten, auszugehen (Biber 1992b: 232) sowie von weniger kontinuierlichen Topiks (d. h. auch von einem hohen Anteil von Switch-Reference, s. 3.8.2):

> [...] it is likely that referential density not only is a property of discourse but also reveals more fundamental cognitive strategies. When reporting an event, speakers must somehow balance their attention between the internal structure of the event (e. g. the particular kind of activity performed) and the participants involved in this event. If referential density is low, this suggests that speakers pay relatively more attention to the event than to the participants; if referential density is high, speakers appear to focus more on the participants. (Bickel 2003: 733)

10 Vgl. Du Bois 1987: 835: „The present corpus happens to include a fair number of relatively short texts, into which the speakers have packed most of the main protagonists of the Pear Film. A different Sacapultec corpus containing longer film narratives, or other genres with lower information pressure, could be expected to show fewer clauses with one lexical argument, and more with zero lexical arguments."

In der Operationalisierung als globaler morphosyntaktischer Parameter über overte Realisierungen von möglichen Argument-Positionen ist die referentielle Dichte allerdings nicht deckungsgleich mit der Informationsdichte als informationsstrukturellem, über den Anteil an Topik-Einführungen berechneten Parameter (s. 3.8.1). Denn die overte Realisierung von Referenten in Argument-Positionen ist zwar ein wichtiges Indiz für die Einführung eines neuen Referenten, aber nicht jede overte Referentenerwähnung muss zwangsläufig ein neues Topik einführen, und nicht jedes neue Topik muss overt realisiert sein (d. h. auch bekannte Referenten können in pro-drop-Sprachen über NPs ausgedrückt werden, ebenso können auch neue Referenten durch Nullmorpheme realisiert werden, vgl. Du Bois 2003: 71; Bickel 2003: 709). So hat beispielsweise das kurze Tiermärchen 1488 im obugrischen Korpus mit 15 Argumentstellen (zwei davon sind Nullanaphern, fünf Pronomen, acht NPs) eine hohe referentielle Dichte (der Anteil overter Realisierung beträgt 13/15, also 0.87). Doch die referentielle Informationsdichte bzgl. der Einführung neuer Referenten ist sehr niedrig (0.09, nur 2 aller 23 Referentenerwähnungen sind Topik-Einführungen, wobei die beiden Hauptreferenten gleichzeitig als Gruppe eingeführt werden). Dagegen sind die obugrischen Texte des Genres der Fate Songs (das sind Personal Songs, s. Kapitel 5) im Korpus gekennzeichnet durch eine mittlere lexikalische referentielle Dichte, da hier der topikale Hauptreferent (der oder die Vortragende) in Subjekt-Position über Nullanaphern realisiert ist (selbst bei Einführung als neues Topik); gleichzeitig haben diese Texte aber durch viele neuen Referenten in adverbialer oder attributiver Funktion (Aufzählung von Landschaftseindrücken) eine hohe allgemeine referentielle Informationsdichte.

Diese Operationalisierung von referentieller Dichte über overte Argumentkodierung kann man – statt als direktes Maß von Informationsdichte – also eher als Hinweis auf die **Explizitheit** der Benennung der Referenten in den zentralen Argument-Positionen verstehen:

> Chinese discourse, for example, is well known for often being very implicit about referents compared to other pro-drop languages such as Spanish. In particular, there is considerable variation in the average ratio of overt argument NPs (nouns or pronouns) to available argument positions in the clause. (Bickel 2003: 708)

In dieser Parametrisierung stellt die referentielle Dichte gewissermaßen das positive Gegenstück zur referentiellen Inferenz dar, die den Anteil nicht overt kodierter referentieller Einheiten angibt (allerdings in abweichender Operationalisierung der Basisanzahl von Referenten über Argument-Positionen).

Wie oben bereits festgestellt, kann auch innerhalb der Sprachpraxis einer Sprechergemeinschaft die referentielle Dichte von Texten mit deren Diskurstyp variieren; sie stellt dementsprechend ein potentielles Merkmal eines TWM als kognitives Genre-Strukturmodell dar (vgl. Schulze 2020: 620, Fußnote 18). Die Dichte expliziter

Referentenerwähnungen kann also ein textkulturspezifischer TWM-Parameter sein, d. h. es kann – wie in obigem Beispiel des Tiermärchens 1488 – ein Merkmal eines Texttyps innerhalb einer Sprechergemeinschaft sein, Referenten explizit zu wiederholen (Wolfgang Schulze, persönliches Gespräch), und damit in der TWM-gesteuerten Text-Modell-Konstruktion der Textverarbeitung den referentiellen Bereich des kognitiven Text-Modells zu betonen (Bickel 2003: 733), vgl. dazu auch Bickel 2003: 732f.: „Referential density does not just depend on random propensities of individuals at given times, but it attests to robust rhetorical norms. […] NP use can […] establish itself as a cultural norm."

3.3.3 Nominal-lexikalische Elaboration

Während sich die beiden eben vorgestellten globalen referentiellen Parameter der referentiellen Dichte und der referentiellen Inferenz allgemein auf die Stärke der Elaboration (bzw. den Inferenzgrad) der Texte bzgl. referentieller Einheiten beziehen (wie viele referentielle Einheiten in einem Text explizit verbalisiert bzw. wie viele Einheiten inferiert werden), so operationalisiert der Parameter der **nominal-lexikalischen Elaboration** die Stärke der qualitativen Elaboration von Referenten durch Attribute (also wie stark die referentiellen Einheiten selbst elaboriert sind).

Diese lexikalische Elaboration referentieller Einheiten wird hier – analog zur Clause-Elaboration – über die durchschnittliche Länge von nominalen lexikalischen Einheiten operationalisiert und bezieht sich damit auf den Grad der **qualitativen Ergänzung** der explizit im Text benannten, lexikalisch als Nominalphrase kodierten Referenten, d. h. der overt kodierten (ohne partiell oder vollständig covert kodierte Referentenerwähnungen, also ohne pronominale Formen oder Nullanaphern; vgl. Schulze 2019: 26f.; 2020: 619). Als TWM-Parameter gibt dieser Parameter dementsprechend an, wie stark die referentiellen Informationseinheiten in einem Text-Modell elaboriert sind. In dieser Arbeit wird diese nominale Elaboration konkret als die durchschnittliche Tokenanzahl pro lexikalischer nominaler Einheit (NP) berechnet:

$$\text{NOM\_ELAB} = \frac{\text{Tokenanzahl lexikalischer nominaler Einheiten}}{\text{Anzahl lexikalischer nominaler Einheiten}} \tag{3.12}$$

3.3.4 Referentielle Explizitheit

Als weiteres Elaborationsmaß referentieller Einheiten kann man über den Anteil der explizit lexikalisch benannten Referenten an den overt (substantiell) benannten Referenten den Grad **lexikalischer referentieller Explizitheit** berechnen (vgl. Schulze 2004a: 556, 561; 2018: 13, 57f.). Basis sind hier – anders als bei der referentiellen Dichte – die über nominale oder pronominale Ausdrücke overt versprachlichten referentiellen Einheiten (ohne Nullanaphern):

$$\text{Ref. Explizitheit} = \frac{\text{Anzahl nominaler Einheiten}}{\text{Anzahl nominaler + pronominaler Einheiten}} \tag{3.13}$$

Informationsstrukturell kann man diesen Parameter des Grades der Explizitheit in der Benennung referentieller Informationseinheiten als Hinweis auf die Bedeutung von anaphorischen Strategien in der Kohärenzherstellung lesen. In der Operationalisierung von Elaboration über lexikalisch kodierte Referentenerwähnungen ist dieser Parameter mit der **referentiellen Dichte** verwandt (s. 3.3.2 bzgl. referentieller Dichte als Explizitheitsmaß), unterscheidet sich aber einerseits durch die Einschränkung auf den Anteil lexikalischer Realisierungen an den overten referentiellen Einheiten (die referentielle Dichte gibt dagegen den Anteil overter, also nominaler und pronominaler Einheiten an allen möglichen Referentenerwähnungen, also inkl. Nullanaphern an); auf der anderen Seite vergrößert die Aufhebung der Beschränkung auf Argument-Position die Grundgesamtheit wieder (also inkl. Adjunkte).

In dieser Arbeit wird eine Operationalisierung der referentiellen Explizitheit über die Tokenanzahl der lexikalisch kodierten nominalen Einheiten gewählt, sodass hier zusätzlich über die Länge der Nominalphrasen der Grad der Explizitheit pro Benennung mitberücksichtigt ist (also deren qualitative Elaboration, s. 3.3.3):

$$\text{REF\_EXPLIC} = \frac{\text{Tokenanzahl lexikalischer nominaler Einheiten}}{\text{Anzahl substantieller referentieller Einheiten}} \tag{3.14}$$

3.4 Globale relationsbezogene Genre-Parameter

Analog zu den globalen referentenbezogenen Parametern können deren relationale Pendants berechnet werden, die sich dementsprechend auf verschiedene Typen der **konzeptuellen Elaboration von Relationen** beziehen.

3.4.1 Relationaler Inferenzgrad

Der **relationale Inferenzgrad** bezieht sich auf den Anteil nicht-overt kodierter relationaler Einheiten (Anteil nicht versprachlichter Relationen, vgl. Schwarz-Friesel & Consten 2014: 70; Schulze 2004a: 561f.), der als Hinweis auf die Stärke der Einbeziehung von relationsbezogenem Weltwissen (vgl. Dik 1997b: 411f.) in den TWM-Aufbau interpretiert werden kann, also als der Anteil der vom Rezipienten zu ergänzenden Relationen. Dieser lässt sich über die relative Häufigkeit von Verbellipsen berechnen (vgl. Martin 2001: 36 bzgl. Ellipse als Kohäsionsmittel):

$$\text{REL\_INFER} = \frac{\text{Vorkommen Verbellipse}}{\text{Anzahl relationaler Einheiten}} \tag{3.15}$$

So wie bei Referenten über typische Argumentstellen können in der Textverarbeitung auch für Relationen über Wissen bzgl. typischer Ereignisfolgen im Sinne von Skripts (s. auch 3.7.2; vgl. 1.1.1) Inferenzen bzgl. nicht versprachlichter Relationen gezogen werden (vgl. Ungerer & Schmid 1996: 213ff.) und diese im kognitiven Modell ergänzt (elaboriert) werden. Der Inferenzgrad eines Textes kann also ein quantitativer Modell-Parameter eines Text-Weltmodells für die Differenzierung von Texttypen sein.

3.4.2 Verbal-lexikalische Elaboration

Im Gegensatz zur relationalen Inferenz, die sich auf den Anteil der in einem Text versprachlichten relationalen Einheiten bezieht, bestimmt der Parameter der **verbal-lexikalischen Elaboration** (vgl. Schulze 2019: 27) die durchschnittliche Länge von verbalen Einheiten (Verbalphrasen) und misst damit, wie stark die explizit verbal kodierten relationalen Einheiten selbst elaboriert sind. In dieser Arbeit wird die verbale Elaboration analog zur nominalen Elaboration berechnet als durchschnittliche Tokenanzahl pro lexikalischer verbaler Einheit (VP):

$$\text{VERB\_ELAB} = \frac{\text{Tokenanzahl der lexikalischen verbalen Einheiten}}{\text{Anzahl lexikalischer verbaler Einheiten}} \tag{3.16}$$

3.4.3 Relationale Explizitheit

Analog zu der referentiellen kann auch die **relationale Explizitheit** als die Stärke der Explizitheit im Ausdruck relationaler Informationseinheiten berechnet werden; diese bezieht sich also darauf, in welchem Umfang die substantiell benannten Relationen explizit verbal benannt werden. Nicht-verbale nominale und adjektivische Einheiten in prädikativer Funktion werden dabei in 6.2.3 als relationale Proformen aufgefasst, gelten hier also als substantielle, aber nicht-explizite relationale Benennung. Wie bei der referentiellen Explizitheit wird auch dieses relationale Explizitheitsmaß in dieser Arbeit über die Tokenlänge der expliziten lexikalischen Einheiten operationalisiert:

$$\text{REL\_EXPLIC} = \frac{\text{Tokenanzahl lexikalischer verbaler Einheiten}}{\text{Anzahl substantieller relationaler Einheiten}} \tag{3.17}$$

3.5 Raum-Zeit-strukturelle Genre-Parameter

Funktionale Genre-Parameter bezüglich der kognitiven Domäne „Raum-Zeit“ beziehen sich auf die **spatiotemporale Strukturierung** des textuell kodierten kognitiven Modells. Dabei kann man die räumliche Struktur eines Text-Modells, die sich auf „Abstandsrelationen bzw. Lagerelation zw. Grenzen, Oberflächen von natürlichen und gemachten Dingen oder Lebewesen“ (Day 2020: 1478) bezieht, als eine

kognitive Karte (*cognitive map*)[11] auffassen (Schulze 2018: 190; 2019: 24), also als „mental model of *spatial* relations" (Ryan 2003: 215, Hervorhebung im Original). Als topologisches Strukturmodell des Text-Modells wird dieses in der Textverarbeitung sukzessive aus den sprachlichen Informationen zur positionellen Relationierung der Textreferenten im Arbeitsgedächtnis aufgebaut. Folgende graphische Darstellung von Schulze 2018: 201 für die kognitive Karte einer udischen Volkserzählung verdeutlicht dieses Konzept:

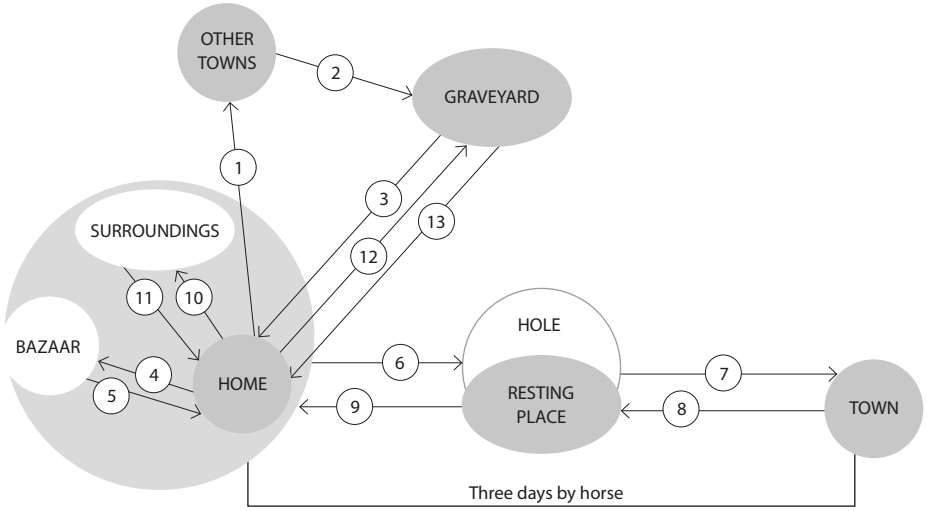


Abbildung 3.2: Kognitive Karte eines udischen Volksmärchens (reproduziert nach Schulze 2018: 201, Abb. 11; vgl. auch Schulze 2020: 618, Abb. 8)

Die schrittweise Konstruktion dieses räumlichen Modells ist dabei mit der (Konstruktion der) zeitlichen Struktur des Text-Modells verknüpft, vgl. Thomaschke 2020: 1954: „[…] zeitliche Ereigniseigenschaften [*werden*] genutzt, um relationale und funktionale Informationen zu gewinnen." Diese **temporale Struktur** bezieht sich also auf die Relationen des Aufeinanderfolgens bzw. der Gleichzeitigkeit von Ereignissen (s. Thomaschke 2020: 1954), die in der Textverarbeitung sukzessive gegeben werden.[12]

11 Mit May 2020 kann eine kognitive Karte definiert werden als „eine räumliche Wissensstruktur, die durch Prozesse der Raumwahrnehmung (primärer Wissenserwerb) oder durch Nutzung von grafischen oder verbalen Raumdarstellungen (sekundärer Wissenserwerb) zustande kommt. Es handelt sich um Wissen im Langzeitgedächtnis (Gedächtnis), das es Menschen, Tieren oder auch Robotern ermöglicht, sich über den aktuell wahrnehmbaren Raumausschnitt hinaus in der Umwelt zu orientieren (Raumorientierung) und nicht sichtbare Ziele zu erreichen (räumliches Navigieren)." (May 2020: 949)

12 Dabei leitet sich die zeitliche Struktur des Text-Modells primär von der Reihenfolge ab, in der die den Text konstituierenden Äußerungen in der Wahrnehmung gegeben sind (s. Schulze 2013). Daneben kann die zeitliche Struktur auch durch sprachliche Temporalisierungsmarkierung spezifiziert werden, etwa durch Kodierung von Vorzeitigkeitsrelationen.

Die kognitive Karte eines Text-Modells dient also dazu, die Positionen bzw. Abstände (Entfernungen) von Referenten im Raum über die Zeit zu verfolgen; durch die schrittweise Aktualisierung der Referentenpositionen im zeitlichen Verlauf ihrer Änderung auf der kognitiven Karte ergibt sich nach und nach die spatiotemporale Struktur des Text-Modells (vgl. Ryan 2003: 215f.; Schulze 2019: 24). Insbesondere werden die topologischen Objektvorstellungen, die als statische Referenzpunkte (Orte) die Elemente der kognitiven Karte bilden, sukzessive in der Verarbeitung der Folge sprachlich kodierter Ereignisvorstellungen konstruiert als der räumliche Hintergrund für die darin (bzw. davor) stattfindenden Ereignisse (Bühnenmodell, vgl. Ryan 2003: 215: „space as a stage for narrative events"). Diese Ortsvorstellungen der kognitiven Karte kann man als *landmarks* auffassen (Langacker 1986; vgl. Schulze 2019: 24), also als zeitstabile, fix positionierte Referenten, relational zu denen die aktiven Referenten als *trajectors* (Langacker 1986) ihre Position verändern: „With the passage of time, one individual, referred to as the **trajector** (*tr*), moves from a position within the neighborhood of another individual, the **landmark** (lm), to a final position outside that neighborhood" (Langacker 1991: 281, Hervorhebung im Original).

So ist etwa für Volksmärchen – im Gegensatz z. B. zu Legenden, die ausgezeichnet sind durch *domain knowledge* (Nguyen u. a. 2012: 380: „Legends are characterized by references to places, persons etc.") – auszugehen von unspezifischen Raum-Zeit-Bezügen (vgl. Schulze 2018: 190; 2020: 607). Die Raum-Zeit-Struktur des durch das entsprechende TWM kodierten narrativen Texttyps zeichnet sich also aus durch „fiktive Landschaft, Raum als stereotype Welt" (Schulze 2020: 607) sowie durch eine temporale Verortung in einer unspezifischen „Vor'-Zeit" (Schulze 2020: 607):

> Typically, the space construed in a folk narrative such as "The Grateful Dead" is grounded in a set of rather generic landmarks that are devoid of being named and hence concretized. The audience can relate these landmarks to their own experience only in the sense that they know of the possible existence of such landmarks. The dimension of such possible spaces is normally confined to the experiential world of the audience, which makes narrative spaces a part of the ecotype of a folk narrative […][.] In other terms: Landmarks in folk narratives reflect highly conventionalized, i. e. socially anchored models of space. (Schulze 2018: 197)

Die Konstruktion der Raum-Zeit-Struktur narrativer kognitiver Modelle in der Textverarbeitung ist entsprechend als **referentenzentriert** anzunehmen – d. h. die Aufmerksamkeit des Rezipienten fokussiert sich auf die Protagonisten und ihre Handlungen (Ryan 2003: 236), die als *trajector* ihre Position innerhalb der kognitiven Karte

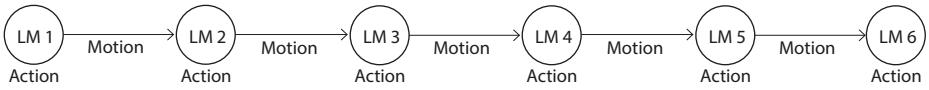


Abbildung 3.3: Landmarks im Zusammenhang mit Handlungs- und Bewegungssequenzen (reproduziert nach Schulze 2018: 198, Abb. 9)

ändern. Die relevanten räumlichen Zusammenhänge werden also nur in das Modell aufgenommen, sofern sie für die Ereignisse der Erzählung als fixe topologische Referenzpunkte (d. h. als *landmarks*) für die Positionsänderung der Hauptreferenten relevant sind (vgl. Schulze 2018: 198, 203; s. auch Abbildung 3.3):

> [...] we can expect that folk narratives usually entail sequences of landmarks, being related by motion acts of the hero. [...] Accordingly, action takes place within the region of the landmarks, whereas the relation between the landmarks is marked mainly for telling the traveling event itself. (Schulze 2018: 197f.)

In der Versprachlichung drücken sich solche spatiotemporalen TWM-Parameter explizit über die Verwendung von Lokal- und Zeitausdrücken, Tempus- und Aspekt-Markern sowie Lokalisierungsmarkern (d. h. Lokalkasus im weiteren Sinn, also auch Adpositionen) aus. Als TWM-Parameter beziehen sich spatiotemporale Modell-Parameter insbesondere auf die Stärke, die Art (etwa Grad der Spezifität) und die sequentielle Verteilung von **Temporalisierungen** und **Lokalisierungen**.[13]

Da die spatiotemporale Struktur von Text-Modellen aus den versprachlichten Ereignissen und Handlungen der Referenten als deren Hintergrund konstruiert wird, vor dem diese stattfinden und der gleichzeitig den Rahmen für die Konzeptualisierung der folgenden Referenten und Ereignisse bereitstellt (vgl. Ryan 2003: 237f.), wird die raum-zeitliche Struktur in dieser Arbeit bei den referentiellen, relationalen sowie informationsstrukturellen Parametern über deren auf die spatiotemporale Strukturierung bezogenen Merkmale mitberücksichtigt.[14] Dies trifft insbesondere auf folgende Parameter zu:

Referentielle Distanz und Topik-Persistenz. Durch eine auf grammatische Relationen bezogene Operationalisierung von referentieller Distanz und Topik-Persistenz können über diese topikalitätsbezogenen Abstandsmaße im adverbialen Bereich (LOC) genrespezifische spatiotemporale TWM-Modellparameter der textuellen Verteilungsmuster von Lokalisierungen und Temporalisierungen abgeschätzt werden (s. 3.6.3).

Ereignisabfolge. Wie oben angedeutet, koinzidieren *landmarks* bei narrativen Textmodellen als Handlungsorte typischerweise mit Abschnitten, in denen die Protagonisten interagieren; sie drücken sich also häufig als Cluster von Handlungsverben

13 So etwa auf die Stärke von Tempusmarkierung als quantitativer Prototyp des schematischen kognitiven Modell-Strukturaufbaus bzgl. der Kategorie Zeit; z. B. drückt sich die zeitlose schematische Struktur in Gesetzestexten durch Fehlen von Tempusmarkierung aus (s. Schulze 2019: 16).

14 Auch die globalen Parameter bzgl. der Elaboration referentieller lexikalischer Einheiten können nach raum- und zeitbezogenen semantischen Domänen differenziert werden und so Hinweise auf die Stärke von spatiotemporalen Informationen im Text-Modell geben (s. Schulze 2019: 27).

aus (Schulze 2019: 24f.), vgl. Schulze 2018: 203: „[…] [*dialogic sequences*] are typically related to events that take place within a landmark. The outcome of the dialogic event lays the ground for the presentation of the action events in the subsequent section." Übergangsbezogene Ereignisfolgemuster der Art `MOTION > ACTION > MOTION`, also Übergänge (>) von einer bewegungsbezogenen Sequenz zu einer Handlungssequenz, die wiederum in eine Bewegungssequenz übergeht (vgl. 3.7.3), können entsprechend als ereignisstruktureller Hinweis auf das Vorliegen von *landmarks* verstanden werden.

Temporal-Sequencing. Sprachabhängig kann Tempusmarkierung die informationsstrukturelle Unterscheidung in Hintergrund-Vordergrund-Informationen mitkodieren (vgl. 3.8.4).

3.6 Referenzfunktionale Genre-Parameter

Im folgenden Abschnitt werden mit den **referenzfunktionalen Parametern** solche potentiellen quantitativen TWM-Merkmale besprochen, die den referentiellen Aufbau eines textuell kodierten kognitiven Modells aus Objektvorstellungen betreffen, also dessen Aktanten-Struktur (vgl. Schulze 2020: 614ff.). Solche quantitativen Operationalisierung messen etwa die Bedeutung eines Referenten im Text-Modell über die Stärke seines Auftretens im Text als Topik im Sinne einer referenzherstellenden sprachlichen Einheit, also über den Grad seiner Diskurstopikalität (s. Chafe 2001: 673f.), vgl. Givón 1983a: 15: „More important discourse topics appear more frequently in the register […]."

Zu diesen quantitativen **Topikalitätsparametern** gehören dabei insbesondere die Anzahl der verschiedenen im Text kodierten Referenten (Referententypes) als Maß absoluter referentieller Informationsdichte, die relative Häufigkeit ihrer Erwähnungen als Grundlage der Identifizierung von Diskurstopiks sowie referentielle Abstandsmaße im Sinne von Givón 1983b als Maßzahlen ihrer syntagmatischen Verteilung im Text (vgl. Schulze 2019: 27).

Die Berechnung dieser quantitativen Referenzstruktur-Parameter, die sprachlich über referentielle Kohärenzmittel wie anaphorische Bezugnahme kodiert werden (vgl. Schwarz-Friesel & Consten 2014: 126), basiert hier in dieser Arbeit (Kapitel 6) auf einer **Referenten-Annotation**, in der auch Referentenerwähnungen durch Nullanaphern in covert kodierten Argument-Positionen beinhaltet sind (s. 5.3.4).

3.6.1 Topikalitätsquotient

Der **Topikalitätsquotient** (*topic quotient*, Thompson 1989: 480; vgl. Cooreman 1987: 3, 211: *degree of topicality*) ist definiert als die relative Häufigkeit der Erwähnungen eines Referenten im Text. Dieser stellt ein basales frequenzbezogenes Topikalitätmaß

dar, das die Relevanz einer referentiellen Informationseinheit in dem zu konstruierenden Text-Modell quantitativ über seine Textfrequenz operationalisiert (s. Myhill 2001: 169),[15] vgl. dazu Thompson 1989: 480: „The topic quotient is meant to measure the global (storywide) importance of an argument."

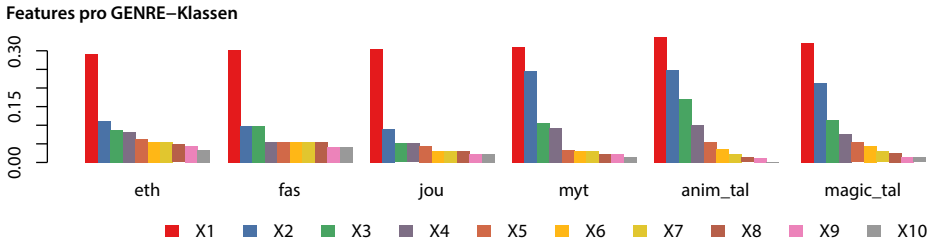
Während bei Thompson (1989) der Topikalitätsquotient über die relative Häufigkeit von Clauses mit entsprechendem Referenten berechnet wird (vgl. Myhill 2001: 169), erfolgt in dieser Arbeit eine automatische Berechnung der relativen Häufigkeit der Referenten im Text anhand der für das obugrische Korpus vorliegenden Daten einer vollständigen Referenten-Annotation. Damit berechnet sich der Topikalitätsquotient eines Referenten X hier wie folgt:

$$\text{TOP\_QUOT(X)} = \frac{\text{Anzahl Erwähnungen von X}}{\text{Anzahl aller Referentenerwähnungen}} \tag{3.18}$$

Die relativen Häufigkeiten der verschiedenen Referenten eines Textes können in einem *bag*-Feature-Set zusammengefasst werden, wobei in dieser Arbeit über eine Frequenzsortierung bzgl. der Referentenhäufigkeit die zehn häufigsten Referenten als Features ausgewählt werden. Über diese Operationalisierung der Verteilung der Topikalität der diskurstopikalsten Referenten kann man dann entsprechende referentielle Strukturtypen feststellen: Für narrative Texte ist von einer geringen Anzahl von Hauptreferenten auszugehen – entsprechend ist eine relativ schwache Topikalität der übrigen Referenten anzunehmen, da sich die meisten Referentenerwähnungen auf wenige Referenten verteilen (vgl. Schulze 2019: 27). Für informative Texte dagegen ist eine eher gleichmäßige Verteilung der Topikalität auf die verschiedenen Referenten anzunehmen.

Wie sich in 6.3.5 bzgl. der Daten des obugrischen Korpus für die entsprechenden Textsorten-spezifischen **Topikalitätsverteilungsmuster** zeigt (s. Plot 3.6.1), gibt es in den untersuchten nicht-narrativen, informativen Texten (etwa `jou`) zwar einen relativ starken Hauptreferenten (`X1`), der – als Thema des Textes – deutlich topikaler ist als die übrigen. Doch zeichnen sich diese nicht-narrativen Texte durch eine viel gleichmäßiger auf die im Topikalitätsranking nachfolgenden Referenten verteilte Topikalitätsstärke aus; die narrativen Texte (etwa `magic_tal`) dagegen haben einige relativ stark topikale Referenten, die Topikalität fällt ab dem zweittopikalsten Referenten entsprechend steiler ab.

15 Vgl. auch Cooreman 1987: 211: „The more frequently a participant occurs in the paragraph, the higher its chance to function as the paragraph theme."



Plot 3.6.1: GENRE-gruppierte Average-Scores-Barplots (Topikalitätsquotienten)

3.6.2 Textweiter Topikalitätsquotient

Bildet man das arithmetische Mittel der Topikalitätsquotienten aller Referenten eines Textes, entspricht dieser textweite durchschnittliche Topikalitätsquotient der inversen Anzahl der verschiedenen Referenten des Texts, also der **inversen Anzahl der Referententypes**:

$$\text{TOP\_QUOT\_TEXT} = \frac{1}{\text{Referententypes}} \tag{3.19}$$

Die textweite Anzahl der unterschiedlichen Referenten (Referententypes) kann man als Genre-differenzierenden Parameter annehmen, der hier als **absolute referentielle Informationsdichte** eines Textes aufgefasst wird (mit dem textweiten Topikalitätsquotienten als dem entsprechenden inversen Maß), vgl. Biber 1992b: 231: „There are two indications of the informational density of a genre here. First, the relative frequency of new references by itself indicates informational focus – as texts introduce more new referents, they increase the informational load."

Für informative Texte ist mit Biber von einer hohen absoluten Informationsdichte auszugehen (d. h. von einem niedrigen textweiten Topikalitätsquotienten, also durchschnittlich schwach topikalen Referenten): „spot news, humanities academic prose, and technical academic prose are extremely 'informational' [...]: they have the highest absolute frequencies of new referents [...]" (Biber 1992b: 232).

Für narrative Texte vergleichbarer Länge[16] ist entsprechend ein niedriger Wert absoluter Informationsdichte anzunehmen, da hier von einem Text-Modell mit wenigen beteiligten Hauptakteuren auszugehen ist, die die Geschichte tragen – im Gegensatz z. B. zu Zeitungstexten, deren kognitives Text-Modell aus einer größeren Anzahl

16 Bei längeren Erzählungen ist von einem größeren Referentenregister auszugehen und entsprechend von einer geringeren durchschnittlichen Topikalität.

an Referenten aufgebaut wird und die entsprechend eine höhere absolute Informationsdichte besitzen.[17]

Dieses (inverse) Informationsdichtemaß wird hier (anders als bei Biber 1992b: 218) nicht bzgl. der Textlänge normiert, da hier der Umfang des Referentenregisters (vgl. Givón 1983a: 10) des Text-Modells (also die Gesamtzahl der in einem Text erwähnten und von der Kognition in dessen Verarbeitung als Elemente des Text-Modells aufzubauenden referentiellen Informationseinheiten) als elementarer quantitativer Modellstrukturparameter anzunehmen ist, der von der Kognition in der Texttypisierung in Text-Weltmodellen als genrespezifischer Wert abgespeichert wird und der in dieser Arbeit entsprechend als TWM-Parameter getestet werden soll. Die in 3.8.1 vorgestellte, relative referentielle Informationsdichte stellt ergänzend als der Anteil neuer (unterschiedlicher) Referenten an den Referentenerwähnungen das entsprechende, bzgl. der referentiellen Textlänge normierte Informationsdichtemaß dar.

3.6.3 Referentielle Distanz

Der Topikalitätsparameter der **referentiellen Distanz** (Givón 1983a) gibt den Abstand koreferentieller Erwähnungen im Verlauf des Textes an; er ist also ein auf die Struktur der sequentiellen Anordnung von Referenten bezogener Parameter. Als solcher wurde die referentielle Distanz von Givón im Rahmen seiner diskursfunktionalen Studien zur Topic-Continuity eingeführt (s. Givón 1983b). Die referentielle Distanz bezieht sich dabei auf den Abstand der Erwähnung einer referentiellen Einheit zu ihrer letzten Erwähnung (also wie lange die letzte Erwähnung zurückliegt; s. Givón 1983a: 13). Die Berechnung erfolgt üblicherweise gemittelt über Konstruktionen und syntaktische oder semantische Funktionen, um diskursfunktionale Abhängigkeiten in der Syntax zu analysieren:

> For each NP in a text, RD [*referential distance*] counts the last time its referent was referred to (including zero anaphora) in the preceding text (e. g. RD = 2 if it was referred to two clauses before) [...]. RD [...] counts make it possible to give a functional profile of a given construction or NP type. [...] This approach has been useful in providing a typological perspective on functional alternations, clarifying the discourse motivations underlying these alternations [...]. (Myhill 2001: 165)

17 Vgl. auch Schulze in seiner Analyse udischer Erzählungen: „Am stärksten ist die Welt der ‚Objekte' elaboriert, wohingegen die Welt der Akteure sich im Wesentlichen auf zwei Protagonisten bezieht. In der Tat ist der Text hochgradig repetitiv, was die Akteure angeht. Die zehn höchstfrequenten Nomina (164 Token) decken 49.8 % aller Nomina ab. Diese Zahl stimmt mit der Annahme überein, dass die TWM von Volkserzählungen auf der Nennung von erwartbaren Einheiten beruhen, was eine entsprechende konzeptuelle Dichte zur Folge hat." (Schulze 2019: 27)

Somit erfasst die referentielle Distanz also die (inverse, da distanzbezogene) **anaphorische Topikalität** im Text: Je kürzer die Letzterwähnung zurückliegt (je niedriger also die referentielle Distanz), desto topikaler ist der Referent an der entsprechenden Textposition (s. Myhill 2001: 165; vgl. auch Givón 1983a: 30). Text- und referenzfunktional kann man den durchschnittlichen Abstand eines Referenten als syntagmatisches Topikalitätsmaß verstehen, das eine Ergänzung zu dem rein auf die Auftretenshäufigkeit von Referenten bezogenen Maß des Topikalitätsquotienten darstellt. So können auch absolut nicht hochfrequente Referenten in einer Textregion gehäuft auftreten (als Diskurs-Subtopiks, vgl. Dik 1997a: 314f.; 1997b: 403) und haben entsprechend (über den gesamten Text hinweg) eine niedrige mittlere referentielle Distanz – ähnlich wie dauerhaft hochfrequente Referenten (etwa Protagonisten in narrativen Texten), und anders als singulär oder verstreut über den Text auftretende Referenten.

Entsprechend hat ein Referent bei seiner Ersterwähnung (d. h. seiner Einführung als neues Topik, das somit noch nicht im Register des Arbeitsgedächtnisses vorhanden ist; vgl. Givón 1983a: 10; s. auch Abschnitt 3.8) einen maximalen Wert referentieller Distanz. Bei der Clause-weisen Abstandszählung bei Givón wird dieser Schwellenwert für **Identifizierbarkeit** bzw. *topic accessibility* (Givón 1983a: 13f.; s. auch 3.8) mit einem Wert von 20 angesetzt (s. Givón 1983a: 13f.; s. auch 6.3.1). In dieser Operationalisierung wird also ein Referent, der in den letzten 20 Clauses nicht erwähnt wurde, als ‚neu' angesehen, also als nicht mehr im aktiven Teil des Arbeitsgedächtnisses vorhanden (s. auch Cooreman 1987: 14).[18]

In dieser Arbeit ist mit der im Korpus vorliegenden Referenten-Annotation (die auch Referentialisierung durch Nullanaphern miteinschließt) eine automatische Berechnung der referentiellen Distanz für jede Referentenerwähnung möglich (statt Clause-weise wie bei Givón); deshalb wird dieser Maximalwert referentieller Distanz für neue oder aus dem Register entfallene Referenten mit 60 angesetzt (d. h. ein Referent wird in dieser Operationalisierung als inaktiv angesehen, sofern er mehr als 60 andere Referentenerwähnungen, inkl. nicht overter Referentialisierungen, zurückliegt). Grund für dieses Vorgehen ist die Annahme von zwei bis drei Referenten in einer prototypischen Ereignisvorstellung (vgl. das Konzept der „Cognitive Transitivity" bei Schulze 2011; s. auch Schulze 2018: 9).

18 Vgl. Cooreman 1987: 14: „Psychologically, the measurement of referential distance roughly measures the speaker's assessment of the ease with which the hearer can identify the referent of a particular argument in the clause. Psychologists have shown that the less recently an item has been mentioned, the harder it is for the hearer to remember and identify this element [...]. We can conceive of fragments of discourse as files in the short term memory in which information at the beginning of a file is gradually lost as new information is added."

```
60-3-1-3
```

Auflistung 3.1: Beispiel für Verlauf referentieller Distanz als Partitur-Folge für Referent X

Die durchschnittliche referentielle Distanz für Referent X berechnet sich allgemein wie folgt (s. 6.3.1):

$$\text{ref-dist} = \frac{\text{Summe (Anzahl Referenten zw. aktueller und letzter Erwähnung von X + 1)}}{\text{Anzahl Erwähnungen von X}} \tag{3.20}$$

Für narrative Texte ist aufgrund von Topic-Continuity-Strategien (vgl. Givón 1983b) auszugehen von hoher Topikalität der Referenten in Subjektposition (so etwa bei Cooreman 1987: 211 für das Chamorro festgestellt;[19] vgl. auch Givón 1983a: 29) und damit von einer niedrigen referentiellen Distanz im Subjektbereich (vgl. Myhill 2001: 165; Givón 1983a: 30). Für den Objekt- sowie den adverbialen Bereich sind relativ dazu entsprechend höhere Werte anzunehmen.

Bei informativen Genres ist dagegen anzunehmen, dass Informationen zu einer Vielzahl verschiedener Referenten gegeben werden (statt zu einigen wenigen Hauptreferenten); entsprechend ist dort von einer höheren referentiellen Distanz im Subjektbereich auszugehen, ebenso von weniger stark ausgeprägten Distanzunterschieden zwischen den verschiedenen Bereichen grammatischer Relationen.

Dass im Obugrischen grammatikalische Strategien zur Aufrechterhaltung von Topic-Continuity angewendet werden, zeigt folgende Kurzauswertung der Daten zur referentiellen Distanz im Korpus obugrischer Texte. In Übereinstimmung mit den Ergebnissen der Studien zur pragmatisch motivierten Objektkodierung im Obugrischen (Nikolaeva 2001; Skribnik 2001; Skribnik 2010; Virtanen 2014; s. auch 5.1.2) kann hier ein Zusammenhang zwischen der Wahl der verbalen Konstruktion und der Topikalität von Agens- bzw. Patiens-Referenten in transitiven Sätzen der obugrischen Sprachen festgestellt werden (s. Abbildung 3.4):

- In Sätzen mit subjektiver Konjugation (als der transitiven Default-Konstruktion) zeigt sich ein im Durchschnitt stark topikales Agens-Argument (niedrige referentielle Distanz), das in diesem Fall die Subjektposition einnimmt.[20]
- Bei Verwendung der objektiven Konjugation (polypersonales Agreement) hat das Patiens-Argument einen im Durchschnitt ähnlich (hohen) Topikalitätsstatus wie

19 Vgl. Cooreman 1987: 211: „I have used a concrete, quantitative method proposed by Givón. The method allowed me to compare the four constructions which occur in basic clauses based on the relative degree of topicality of Agents and Objects in the transitive propositions. The application of the quantitative method on the Chamorro narratives lead to the following general observations: 1. The highest topical NP in a proposition is selected to fulfill the role of syntactic subject in the sentence."

20 Im Obugrischen gibt es sowohl monopersonales also auch polypersonales Agreement, die sog. *subjektive* sowie die *objektive* Konjugation; s. dazu 5.1.2.

das Agens-Argument (es ist sog. *secondary topic*, s. Nikolaeva 2001); beide haben also einen niedrigen Wert referentieller Distanz.

- In Sätzen mit Passivkonstruktion (d. h. mit Demotion des Agens-Arguments aus der Subjektposition) ist dagegen das Patiens-Argument, das hier im Passiv entsprechend die Subjektposition einnimmt, im Durchschnitt topikaler (niedrige durchschnittliche referentielle Distanzwerte).

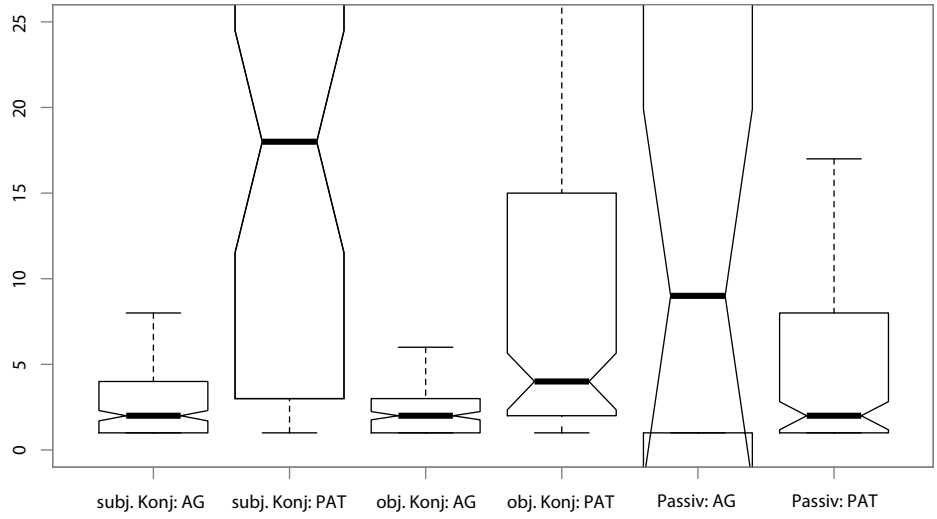


Abbildung 3.4: Durchschnittliche referentielle Distanz von Agens- und Patiens-Argument in transitiven Sätzen des obugrischen Korpus (differenziert nach Agreement-Typen bzw. Diathese)

3.6.4 Topik-Persistenz

Das ebenfalls von Givón etablierte Maß der **Topik-Persistenz** misst die Anzahl verbleibender Erwähnungen eines Referenten (also wie oft der Referent sich danach noch im Text wiederholt; s. Givón 1983a: 14f.). Als spiegelbildlicher Parameter zur referentiellen Distanz erfasst dieses Maß die **kataphorische Topikalität**:

> [...] TP [*topic persistence*] counts how many times it is referred to in the following text (e. g. TP = 1 if it is referred to again in the following clause but not in the clause after that). We can say that an NP is generally more topical if its RD is low and its TP is high, but of course we are really measuring two types of topicality here, anaphoric (RD) and cataphoric (TP). (Myhill 2001: 165)

Als Parameter der kognitiven Textverarbeitung misst die Topik-Persistenz, wie lange sich ein Referent im Prozess der Konstruktion eines Text-Modells im Arbeitsgedächtnis hält, wie lange er also im Referentenregister des sukzessive aufgebauten kognitiven Modells verbleibt (vgl. Givón 1983a: 15). Dieser Parameter bezieht sich damit

auf den Grad der vorwärtsgewandten Fortsetzung eines Referenten im Gedächtnis, im Gegensatz zur referentiellen Distanz, die ein Maß für den rückwärtsgewandten Verlauf ist (s. Abschnitt 3.8 zur Aktivierung und Re-Aktivierung von Referenten als informationsstrukturellen Größen; vgl. auch Cooreman 1987: 15).[21]

Interpretieren lässt sich der Parameter der Topik-Persistenz, der die regionale Verteilung der Stärke von Referenten abbilden kann (also in welchem Teil des Textes sich ein Referent wie stark fortsetzt), als Relevanz eines Referenten als Diskurstopik (s. Givón 1983a: 14), vgl. Givón 1983a: 15: „More important discourse topics appear more frequently in the register, i. e. they have a higher probability of persisting longer in the register after a relevant measuring point."

Berechnet als globaler textfunktionaler Wert, also über textweite Durchschnittsbildung pro Referent (d. h. ohne regionale Differenzierung), ist dieser Wert ähnlich wie der Topikalitätsquotient ein textfrequenzbezogenes Maß der Topikalitätsstärke, allerdings in anderer Skalierung bzw. Zentrierung: Die Topik-Persistenz eines Referenten ist bei singulärer Erwähnung 0 und (im Gegensatz zum Topikalitätsquotienten) ein absolutes Maß der Wiederholung eines Referenten im Text oder einer Textregion (nicht relativ zu Gesamtzahl der Referenten). Die Topik-Persistenz ist also ein Maß für die Stärke vorwärtsgewandter Topikalität (vgl. Givón 1983a: 30).

Mit der hier im Korpus vorliegenden Referenten-Annotation ist eine automatische Berechnung der Topik-Persistenz für jede Referentenerwähnung möglich (statt Clause-weise wie bei Givón):

```
3-2-1-0
```

Auflistung 3.2: Beispiel für Topik-Persistenz-Verlauf als Partitur-Folge für Referent X

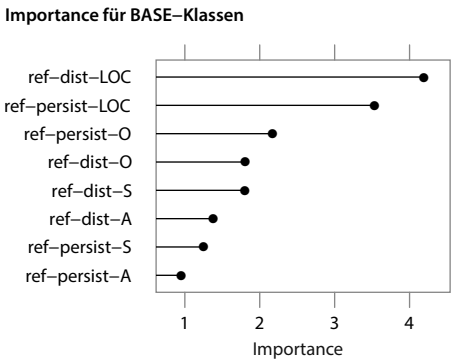
Die durchschnittliche Topik-Persistenz für Referent X (ggf. in einer spezifischen Region R) berechnet sich damit wie folgt (Auflösung mit Gaußscher Summenformel; Details s. 6.3.2):

$$\text{ref-persist} = \frac{\text{Summenreihe 1 bis (Anzahl verbleibender Erwähnungen von X in R) - 1}}{\text{Anzahl Erwähnungen von X in R (= n)}} = \frac{n - 1}{2} \tag{3.21}$$

Für die Typik narrativer Texte ist bzgl. der Topik-Persistenz – indirekt proportional zur Typik referentieller Distanz – bei Vorliegen von Topic-Continuity-Strategien (vgl. Givón 1983b) von einem topikalen Referenten in Subjektposition auszugehen (vgl. Cooreman 1987: 211) und damit entsprechend von einer hohen Topik-

21 Vgl. Cooreman 1987: 15: „The parameter of persistence is roughly related to the speaker's intention in language production, i. e., the way (s)he plans ahead which entities should be important and thus continued as topics in the narrative sequel. One would expect that important topics occur frequently in the text so that they tend to show a relatively low value for referential distance and a relatively high one for topic persistence."

Persistenz im Subjektbereich. Für den Objekt- sowie den abdverbialen Bereich sind relativ dazu entsprechend niedrigere Werte anzunehmen.



Plot 3.6.2: Feature-Importance für BASE-Klassen (Referentielle Distanz und Topik-Persistenz)

Mit den beiden abstandsbezogenen Topikalitätsparametern der referentiellen Distanz und der Topik-Persistenz lassen sich in einer Operationalisierung mit Durchschnittsbildung bzgl. syntaktischer Rollen (grammatischer Relationen; vgl. Schulze 2004a: 562) z. B. ortsbezogene TWM-Modelle von solchen mit unspezifischer Lokalisierung wie in Volksmärchen (s. Abschnitt 3.5) differenzieren. So zeigen sich etwa im obugrischen Korpus die referentielle Distanz und die Topik-Persistenz im lokativischen Bereich als die wichtigsten Textsorten-differenzierenden Distanz- bzw. Persistenz-Features, s. den Feature-Importance-Plot 3.6.2.

3.7 Relationsfunktionale Genre-Parameter

Im folgenden Abschnitt werden quantitative **relationsfunktionale Parameter** besprochen, die den Aufbau der **Ereignis-Struktur** eines textuell kodierten kognitiven Modells betreffen. Da sich Ereignisvorstellungen über relationale Einheiten ausdrücken (vgl. Schulze 2018: 59: „[…] relators can be seen as the meronymic expression of generalized event images"), basiert die Berechnung der Parameter dieser relationssemantischen Textstruktur auf quantitativen Maßen zu den die Relationierungsfunktion gewährleistenden Einheiten; es handelt sich hier also um quantitative Modelle von **Prädikaten**.

Insbesondere sind hier die Frequenzverteilung und (häufige) Abfolgen von Typen verbaler Relationen im Text relevant (Ereignistypik und Ereignisabfolgen, vgl. Schulze 2020: 619ff.). Von diesen relationalen Ereignis-Parametern kann angenommen werden, dass sie wesentliche Elemente eines TWM-Modells darstellen:

> Neben der Beschreibung des Elaborationsgrades von referentiellen Einheiten ist die Darstellung von Typen der Ereignisdarstellung und Ereignisabfolgen sicherlich ein zentrales Moment, um die Genre-Grammatik eines narrativen Textes greifbar zu machen. (Schulze 2020: 619)

Im Gegensatz zu den referentiellen Einheiten, für die Frequenzen für jede einzelne dieser Informationseinheiten berechnet werden können, werden bei den Relationen deren semantische Typen (Verbklassen, vgl. Schulze 2019: 27f.) als **Ereignistypen**

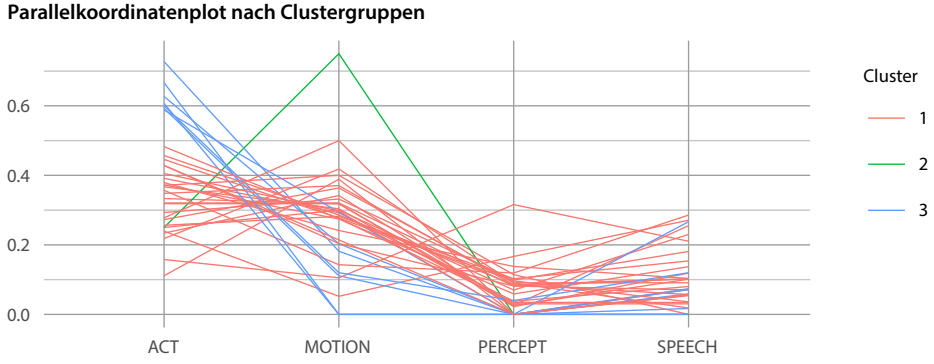
(etwa Handlung, Bewegung, Wahrnehmung, Zustand) Hauptgegenstand der TWM-Parametrisierung sein. Denn in einem Text ist zwar immer nur eine begrenzte Anzahl an Referenten kodiert (also an handelnden Personen, beteiligten Objekten, Schauplätzen der Handlung usw.), deren Textfrequenzen entsprechend als TWM-Parameter dienen können – dagegen kann aber (je nach Textumfang) eine hohe Anzahl verschiedener, jeweils neuer Ereignisse (als Relationen zwischen diesen Referenten) im Text versprachlicht werden, sodass für die Operationalisierung der schematischen Ereignisstruktur von Texten entsprechend auf Ereignistypen Bezug genommen wird.

3.7.1 Ereignistypik

Über die relative Textfrequenz von semantischen Verbklassen kann die Stärke der verschiedenen Ereignistypen im Text berechnet werden und somit ein quantitatives Textprofil der **Ereignistypik** erstellt werden.

Ereignistypen:	ACT	MOTION	PERCEPT	SPEECH	STATE
relative Häufigkeit der Ereignistypen:	0.4	0.3	0.1	0	0.2

Tabelle 3.4: Beispiel einer Frequenzverteilung für Ereignistypen



Plot 3.7.1: Parallelkoordinatenplot nach Clustergruppen (Ereignistypik)

Für die Ereignistypik von Volkserzählungen ist von einer Dominanz von Ereignisvorstellungen des Handlungs- und Bewegungsbereichs auszugehen (vgl. Abschnitt 3.5; vgl. Schulze 2019: 21, 27f.). Allgemein ist anzunehmen, dass die Ereignistypik ein wichtiger Genre-differenzierender Parameter ist (vgl. Schulze 2020: 619f.); vgl. dazu die Cluster-spezifische Auswertung der Ereignistypik in Plot 3.7.1 für das obugrische Korpus mit einem hohen Anteil von Volkserzählungen, die eine Hauptgruppe (rot) mit hohem `ACTION`- und `MOTION`-Anteil sowie davon abweichende Typen zeigt.

3.7.2 Ereignisabfolge

Neben der Ereignistypik kann auch die **Ereignisabfolge** als die lineare Abfolgestruktur der im Text kodierten Sequenz von Ereignissen (s. Schulze 2019: 21, 24f.; vgl. Ungerer & Schmid 1996: 214: *event sequences*) über globale Muster von Verbklasse-Folgen berechnet werden. Diese Ereignis-Sequenzmuster erlauben u. a. Rückschlüsse auf die einem narrativen Texttyp zugrunde liegenden, globalen strukturellen Handlungsablauf-Schemata im Sinne von Propp 1972 (vgl. 2.1.2). Solche textsortentypischen Strukturmuster können dann über die Auswertung von Texten als Folgen von Verbklassen mit den in Abschnitt 4.4 vorgestellten Extraktions- und Klassifizierungsmethoden für Sequenzen untersucht werden:

```
STATE-ACT-ACT-MOTION-ACT-...
```

Auflistung 3.3: Beispiel einer Folge von Ereignistypen (Sequenz der Verbklassen)

Für Text-Weltmodelle von Volkserzählungen kann als prototypische Ereignisabfolge, die für narrative Texttypen allgemein als chronologische Kette von Handlungen von Referenten zu bestimmen ist (s. 5.2.4; vgl. de Beaugrande & Dressler 1981: 190f.; Heinemann & Viehweger 1991: 237), nach dem Frame-Setting (häufig mit Zustandsverben wie z. B. bei der Einleitung *Es war einmal*) ein Wechsel von Blöcken von Handlungs- und Bewegungsverben angenommen werden (vgl. Schulze 2020: 607; vgl. auch Abschnitt 3.5 bzgl. *trajector-landmark*-Folgen als kognitive spatiotemporale Grundstruktur von Erzählungen). In einer Operationalisierung über Sequenzen von **Übergängen** von Verbklassetypen, die von der Verweildauer in den Ereignistyp-Zuständen absehen (vgl. auch 3.7.3) und über die man (durch die Einschränkung auf Übergänge von Ereignistypen) ein übersichtliches relationales Profil der Handlungsstruktur eines Textes erstellen kann, stellt sich dieses prototypische Muster z. B. folgendermaßen dar:

```
(STATE)-1-(STATE>ACT)-2-(ACT>MOTION)-1-(MOTION>ACT)-...
```

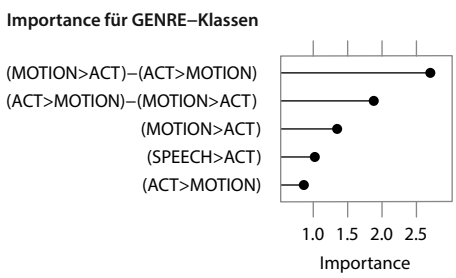
Auflistung 3.4: Beispiel einer Folge von Übergängen zwischen Ereignistyp-Zuständen (Sequenz der Wechsel der Verbklassen)

3.7.3 Häufige Ereignisabfolge-Muster

Neben der auf die globale Handlungsstruktur bezogenen Operationalisierung textweiter Abfolgen von Ereignistypen können über **Frequent-Patterns** von Verbklassen auch die für Texte eines Genres typischen **lokalen** Teilmuster von Ereignisabfolgen als TWM-Parameter berücksichtigt werden.

Kognitiv kann man solche lokalen Ereignisabfolge-Muster als strukturelle **Chunks** interpretieren, die durch Reduktion (Zusammenfassung einer spezifischen Folge von

Einheiten zu einer komplexen Einheit) die Verarbeitung großer Informationsmengen durch das Arbeitsgedächtnis trotz dessen beschränkter Kapazität ermöglichen (s. 1.1.1, Fußnote 3; vgl. auch de Beaugrande & Dressler 1981: 95f.). Solche schematischen Wissensstrukturen typischer, also wiederholter Ereignisabfolgen sind auch Grundlage für Inferenzschlüsse, also für die Ergänzung evtl. fehlender Ereignisabfolgen (s. 3.4.1; vgl. Ungerer & Schmid 1996: 213ff.), vgl. auch de Beaugrande & Dressler 1981: 95 (Hervorhebung im Original): „SCHEMATA sind globale Muster von Ereignissen und Zuständen in geordneten Abfolgen, wobei die Hauptverbindungen in zeitlicher Nähe und Kausalität bestehen. […] Anders als ein ‚Rahmen' ist ein Schema immer als Reihenfolge so aufgestellt, daß Hypothesen gebildet werden können, was in einer Textwelt als nächstes getan oder erwähnt werden wird."



Plot 3.7.2: Feature-Importance für GENRE-Klassen (Häufige Ereignisübergänge)

Für Volkserzählungen kann man annehmen, dass diese geprägt sind durch sich abwechselnde Folgen von Bewegungs- und Handlungssequenzen unterschiedlicher Länge (vgl. Abschnitt 3.5), also durch typische Muster von Ereignistyp-Blöcken der Art `MOTION > ACTION > MOTION` (s. Schulze 2020: 607: „Lexikalische Ausdrücke für Motion und ‚Handlung' bilden tendenziell jeweils Cluster"). So zeigt sich auch in der Feature-Importance-Auswertung für häufige Ereignisabfolgen des obugrischen Korpus, dass die Präsenz ebensolcher Ereignisfolgen in den narrativen Texten für die Differenzierung zwischen Volksmärchen und den anderen Textsorten zentral ist (s. Plot 3.7.2; vgl. auch Report 6.4.2).

Der Wechsel zwischen den Blöcken von Ereignistypen kann formal als Übergang (*transition*) von einem Ereignistyp-Zustand in einen anderen beschrieben werden (s. 6.1.2.6). Operationalisiert werden können diese häufigen Ereignistyp-Muster dementsprechend über die Analyse der relativen Häufigkeit bzw. der Präsenz von N-Grammen solcher Verbklassen-Übergänge (vgl. Tabelle 3.5).

Ereignistyp-Übergänge:	(SPEECH>ACT)	(MOTION>ACT)	(MOTION>ACT)-(ACT>MOTION)
Präsenz des Musters:	0	1	1
relative Häufigkeit des Musters:	0	0.09	0.06

Tabelle 3.5: Beispiel für Frequent-Patterns von Ereignistyp-Übergängen

3.8 Informationsstrukturelle Genre-Parameter

Als letzter funktionaler Parametertyp werden solche Parameter behandelt, die sich auf die Struktur der **Informationsübertragung** beim Aufbau des textuell kodierten

kognitiven Modells beziehen (vgl. Schulze 2019: 28ff.; 2020: 623ff.). Wurde mit den zeit-räumlichen, referentiellen und relationalen Parametern bisher nach der funktional-inhaltlichen Struktur gefragt – d. h. welche Informationen textuell kodiert werden, wie viele Informationseinheiten auftreten, welcher Art diese sind, wie sie sich über den Text verteilen –, so wird jetzt die funktional-informationelle Struktur untersucht, also wie diese Informationen übermittelt und kodiert werden.[22] Vgl. auch Skopeteas u. a. 2006 zum Konzept des *information packaging* bei Chafe (1976):

> [...] Chafe (1976) speaks about 'information packaging' and considers hypotheses about the receiver's assumptions as crucial to discourse structure. These are hypotheses about the status of the referent of each linguistic expression, as represented in the mind of the receiver at the moment of utterance. Thus it is the way the information is transmitted that is crucial, rather than the lexical or propositional content of a sentence, around which grammar usually centers. (Skopeteas u. a. 2006: 1)

Basierend auf den in 1.1.1 getroffenen Feststellungen über die kognitiven Konstruktionsprozesse des sukzessiven Aufbaus mentaler Text-Modelle im Arbeitsgedächtnis bei der Verarbeitung von Texten als Äußerungsfolgen, lassen sich mit dem *gedächtnisbezogenen* **Aktivierungsstatus** sowie dem *bewusstseinsbezogenen* **Aufmerksamkeitsstatus** zwei grundlegende informationsstrukturelle Kategorien unterscheiden (s. Skopeteas u. a. 2006: 2). Diese beiden Kategorien geben eine entsprechende Systematik für die in dieser Arbeit relevanten informationsstrukturellen Parameter vor, die sich gleichermaßen auf referentielle wie auch auf relationale Informationseinheiten beziehen lässt. Die Darstellung orientiert sich dabei primär an Skopeteas u. a. 2006, die (mit Bezug auf Lambrecht 1994 und Chafe 1976) folgende Festlegungen für eine Aktivierungsskala treffen:[23]

> [...] an active concept is given (it is then a topic), and an inactive one is new. [...]
>
> - active concept: one that is currently lit up, a concept in a person's focus of consciousness at a particular moment.
> - semi-active (accessible) concept: one that is in a person's peripheral consciousness, background consciousness
> - inactive concept: one that is in a person's long-term memory, neither focally nor peripherally active.
>
> (Skopeteas u. a. 2006: 2)

22 Vgl. Schwarz-Friesel & Consten 2014: 105: „Wie entfaltet sich die Information im Text? Wie ist das Verhältnis von bekannter, alter und unbekannter, neuer Information und entsprechend die Wechselwirkung von Aktivierung, Re-Aktivierung und De-Aktivierung?"

23 Die bei Skopeteas u. a. 2006 entwickelte Systematik sprachlicher Parameter für informationsstrukturelle Fragestellungen ist auch Grundlage der sprachtypologischen Darstellung der obugrischen Sprachen in 5.1.2.

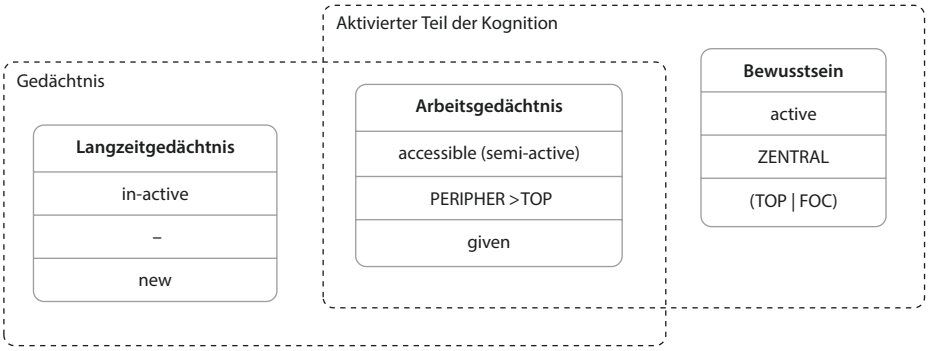


Abbildung 3.5: Kognitiver Aktivierungsstatus in Gedächtnis und Bewusstsein

Demnach (Skopeteas u. a. 2006: 2f.; Lambrecht 1994: 76, 88, 93ff.; vgl. auch Chafe 1976; 1987) kommt es durch Äußerung eines sprachlichen Ausdrucks in einer Kommunikationssituation zu einer Aktivierung (Bewusstmachung) der mit diesem verknüpften, im Langzeitgedächtnis abgespeicherten Informationseinheit (Konzept); diese bisher inaktive (in der Kommunikationssituation noch nicht vorerwähnte, also neue) Einheit wird also aus dem Langzeitgedächtnis abgerufen und damit aktiviert (Topik-Einführung, s. 3.8.1).

Wenn die Information im Moment der Verarbeitung durch die Aufmerksamkeitssteuerung des Bewusstseins als für das Text-Modell relevant (zentral) markiert ist (Fokus, s. 3.8.3; vgl. auch 3.8.4 bzgl. Vordergrund-Hintergrund-Struktur), wird diese im Text-Modell im Arbeitsgedächtnis als dem aktivierten Teil des Gedächtnisses (Dreisbach 2020: 480) abgespeichert, in dem das kognitive Modell des Textes gemäß des schematischen TWM-Strukturmodells seines Genres aufgebaut wird.

Dort steht diese Informationseinheit dann als **peripher** bewusstes (semi-aktives) Element des Text-Modells zur Wiederaufnahme (Re-Aktivierung) zur Verfügung; sie ist also für das Bewusstsein zugänglich (*accessible*). Als bekannte Information (*given*) wird diese typischerweise als Satz-Topik versprachlicht (Perspektivierung, s. 3.8.2).

3.8.1 Informationsfluss und -dichte (Topik-Einführung)

Der Parameter der **Informationsdichte** (Biber 1992b; Du Bois 1987; 2003; vgl. 3.3.2) bzw. des **Informationsflusses** (*information flow*: Chafe 1987; 2001; vgl. Schulze 2020: 623ff.) bezieht sich auf den Umfang bzw. auf die sequentielle Verteilung der Aktivierung referentieller Informationseinheiten aus dem Langzeitgedächtnis, die in der Textverarbeitung durch die Erwähnung von Referenten in einem Text ausgelöst wird. Diese neu in den Diskurs eingeführten, konzeptuellen Einheiten (*new topics*, s. Dik 1997a: 213; Schulze 2018: 73) kann man mit Schulze 2018 auffassen als „Cognitive Topic" im Sinne eines satzübergreifenden „topic as a function referring to the emergence of knowledge states in sequences of narrated scenarios" (Schulze 2018: 73).

Bei einer **Topik-Einführung** wird diese mit dem referenzierenden Ausdruck verknüpfte referentielle Informationseinheit aus dem Langzeitgedächtnis abgerufen (aktiviert), gelangt dabei in das Bewusstsein und kann als dann zugängliche, bekannte Information in das Text-Modell integriert werden, das als kognitives Modell schrittweise im Arbeitsgedächtnis (dem Speicher des aktivierten Teils der Kognition) aufgebaut und durch jede neue Äußerung um die im Text kodierten zentralen Referenten und Relationen erweitert wird. Entsprechend sagt der gedächtnisbezogene Aktivierungsstatus eines Referenten bei seiner Erwähnung aus, ob dieser zu dem Zeitpunkt bereits im aktivierten Bereich des Gedächtnisses vorhanden (***given***), also als Teil des kognitiven Text-Modells für das Bewusstsein zugänglich ist (*accessible*, s. *topic accessibility* bei Givón 1983a: 17ff.; vgl. auch Myhill 2001: 165), oder ob der Referent erstmals erwähnt und entsprechend als neue Informationseinheit in das Text-Modell integriert werden muss (***new***):

> Aus kognitiver Perspektive werden Thema und Rhema als Informations- oder Aktivierungszustände im Arbeitsgedächtnis des Lesers betrachtet. Übertragen wir dies auf die Konzeption des Textweltmodells, dann haben thematische Einheiten im Textweltmodell [*hier im Sinne eines konkreten kognitiven Text-Modells*] bereits eine Repräsentationseinheit und re-aktivieren diese nur, während rhematische Textelemente neue Einheiten im Textweltmodell etablieren oder erstmalig aktivieren. (Schwarz-Friesel & Consten 2014: 105)

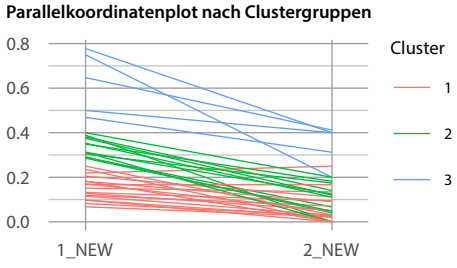
Das zentrale Kriterium zur Feststellung des gedächtnisbezogenen Aktivierungsstatus von Referenten ist also das Vorhandensein im kognitiven Text-Modell, genauer im assoziierten mentalen Register; gegebene Referenten sind also „[…] topics which the speaker assumes the hearer can identify uniquely, is familiar with, are within his file (or register) and thus available for quick retrieval“ (Givón 1983a: 10). Dabei können die Referenten sowohl textuell als auch kontextuell gegeben, d. h. für das Bewusstsein zugänglich sein (s. Skopeteas u. a. 2006: 2). In der in dieser Arbeit angestrebten korpuslinguistischen Operationalisierung kann der Aktivierungsstatus nur über die **Text-Givenness** bestimmt werden, sodass hier gilt: „[…] a given constituent is one which is entailed by the preceding discourse. This use of givenness is of course restricted to text-givenness, as opposed to context-givenness“ (Skopeteas u. a. 2006: 2). In dieser Operationalisierung gilt also entsprechend jede Ersterwähnung eines Referenten im Text als Einführung eines neuen Topiks (vgl. Biber 1992b: 220).

Es ist anzunehmen, dass die Anzahl und die sequentielle Verteilung der Einführung neuer Topiks im Text ein wichtiges Merkmal für die schematischen TWM-Strukturmodelle des typischen informationsstrukturellen Aufbaus kognitiver Text-Modelle sind, vgl. Schulze 2020: 623: „Ein wesentliches Moment bei der Verarbeitung von Volkserzählungen ist die Frage, in welchem Umfang sich die Hörer auf ‚Neues‘ einlassen müssen […].“ Über den Anteil neuer Topiks an der Gesamtzahl an

Referentenerwähnungen im Text, operationalisiert als die relative Frequenz textueller Ersterwähnungen, kann dessen **relative referentielle Informationsdichte** berechnet werden, vgl. Biber 1992b: 231f.: „the relative proportions of new and given references indicate relative informational focus."

$$\text{Informationsdichte} = \frac{\text{Anzahl neuer Topiks (Ersterwähnungen)}}{\text{Anzahl der Referentenerwähnungen}} = \frac{\text{Referententypes}}{\text{Referententokens}} \tag{3.22}$$

Mit Biber (1992b: 231f.) kann man davon ausgehen, dass sich die Informationsdichte je nach Genre unterscheidet: Als potentieller TWM-Parameter ist für die relative referentielle Informationsdichte narrativer Texte zu erwarten, dass diese im Vergleich mit informativ-expositorischen Texten bedeutend niedriger liegt, dass also etwa in Volkserzählungen durchschnittlich weniger neue Topiks eingeführt werden als z. B. in Nachrichtentexten: „It is noteworthy that spot news, [...] and [...] academic prose are extremely 'informational' in both respects: they have the highest absolute frequencies of new referents, and proportionally they use very high percentages of their referring expressions for new references" (Biber 1992b: 232). So eine niedrige referentielle Informationsdichte, wie sie für narrative Text anzunehmen ist, kann man gleichzeitig aber auch als hohe relationale Informationsdichte interpretieren, da entsprechend mehr neue Informationen über die einzelnen Referenten gegeben werden; vgl. dazu Bickel 2003: 733: "If referential density is low, this suggests that speakers pay relatively more attention to the event than to the participants [...]."



Plot 3.8.1: Regionale Verteilung nach Clustergruppen (Topik-Einführungen)

Region 1	Region 2
0.4	0.1

Tabelle 3.6: Beispiel einer regionalen Verteilung von Topik-Einführungen (relative Häufigkeit neuer Referenten pro Region)

Über den Parameter der Informationsdichte kann auch der **Informationsfluss** operationalisiert werden, also der sequentielle Verlauf der Einführung neuer Referenten im Text (vgl. Schulze 2020: 623ff.). Dazu wird in dieser Arbeit die regionale Verteilung der Informationsdichte in Textpartitionierungen bestimmt, also der Anteil neuer Topiks in einzelnen Textregionen, insbesondere für die erste und zweite Texthälfte (s. `1_NEW` und `2_NEW` in Plot 3.8.1).

Durch die Operationalisierung in einem Feature-Set der Informationsdichte in verschiedenen Textregionen können Informationsdichte und -fluss eines Textes gemeinsam ausgewertet werden. Für die erste Region (bzw. bei nicht regional differenzierter Berechnung der textweiten relativen Informationsdichte) entspricht der Wert der Stärke neuer Topiks durch

deren Definition über die Text-Givenness dem Type-Token-Verhältnis der Referentenerwähnungen, denn in diesem Fall ist die Anzahl neuer Topiks gleich der Anzahl der verschiedenen Referenten (Referententypes) im Abschnitt; für die weiteren Textregionen gilt dies nicht mehr, da die dort vorkommenden Referententypen schon in einer vorhergehenden Region eingeführt worden sein können (diese können also statt *new* bei ihrer Ersterwähnung ebenso *given* sein).

$$\text{Informationsdichte erster Region (1\_NEW)} = \frac{\text{Referententypes}}{\text{Referententokens}} \tag{3.23}$$

Bei einer binären Partitionierung kann als Kennzahl für den Informationsfluss auch der Grad der Abnahme der Einführung neuer Referenten zwischen den beiden Textregionen über das Verhältnis der regionalen Informationsdichten berechnet werden:

$$\text{Abnahme der Informationsdichte (bei 2 Regionen)} = 1 - \frac{\text{2\_NEW}}{\text{1\_NEW}} \tag{3.24}$$

Auch für den textuellen Informationsfluss kann man davon ausgehen, dass sich dieser je nach Genre unterscheidet (s. Schulze 2020: 607, 623), dass also die grundlegende referentielle Informationsstrukturierung eines Textes – sprachlich primär über Anaphorik, Wortstellung, Kasus und Agreement kodiert – mit der Kommunikationssituation variiert (vgl. van Dijk 2018: 32). Für den Informationsfluss von Volkserzählungen ist mit Schulze „nach Einführung von Topiks [*eine*] geringe Anzahl an ‚New Topics'" (Schulze 2020: 607) anzunehmen, also eine deutliche Abnahme der Informationsdichte im Verlauf des Textes. Für informative Texte ist dagegen eine insgesamt höhere Informationsdichte sowie eine weniger starke Abnahme der Informationsdichte im Verlauf des Textes anzunehmen.

Wie man in der gemeinsamen Auswertung der Feature-bezogenen TWM-Parameter in 6.6.1 sieht (s. insbesondere Plot 6.6.4), ergänzen sich der in 3.6.2 als Topikalitätsmaß eingeführte Parameter des textweiten durchschnittlichen **Topikalitätsquotienten**, der sich als inverses Maß **absoluter Informationsdichte** (Anzahl der Referententypes) auf den spezifischen referentiellen Umfang eines kognitiven Text-Modells bezieht und entsprechend Text-Modelle unterschiedlichen referentiellen Umfangs differenziert (etwa narrative Subgenres wie kurze Tiermärchen von längeren Erzählungen), und die hier vorgestellte relative referentielle Informationsdichte (vgl. Biber 1992b: 231f.), die als relativer Anteil neuer Topiks an der Anzahl der Referentenerwähnungen eines Textes einen von der Textlänge eher unabhängigen Wert für die Strukturierungstypik bzgl. der durchschnittlichen Frequenzrate der Präsentation neuer referentieller Informationen angibt (der für narrative Texte unabhängig von ihrer Länge ähnlich niedrig anzunehmen ist; vgl. hierzu Tabelle 3.7).

Texttyp	Referententokens (referent. Textlänge)	Referententypes (absolute Dichte)	1/Referententypes (Topikalitätsquotient)	Referententypes/-tokens (relative Dichte)
informativ	10	4 (+)	0.25 (-)	0.4 (+)
narrativ lang	20	4 (+)	0.25 (-)	0.2 (-)
narrativ kurz	10	2 (-)	0.5 (+)	0.2 (-)

Tabelle 3.7: Absolute und relative Informationsdichte (fiktive Beispieldaten zur erwartbaren Differenz zwischen informativen Texten sowie Erzähltexten umfangreicher bzw. kurzer Länge)

3.8.2 Perspektivierung (Switch-Reference-Struktur)

Der Parameter der **Perspektivierung** bezieht sich darauf, welcher Referent im Zentrum der Aufmerksamkeit steht (bzw. ob Referenten nach ihrer Einführung oder Wiederaufnahme im Zentrum bleiben), also über welchen Referenten in einer Folge von Äußerungen jeweils neue (oder kontrastive) Informationen gegeben werden. Dieses Satz-Topik[24] wird typischerweise in **zentraler** syntaktischer Position (Subjekt) kodiert (vgl. Schulze 2018: 79); entsprechend wird der Parameter der Perspektivierung über den Wechsel des Referenten in der Subjektposition operationalisiert:

> Ein Textreferent [...] ist salient, während er im KZG [*Kurzzeitgedächtnis, hier: Bewusstsein*] verarbeitet wird, er ist dann im Aufmerksamkeitsfokus. Ob er über längere Passagen der Textrezeption salient bleibt oder als Nebenfigur bald wieder de-aktivert [*sic*] wird, hängt von Texteigenschaften auf zwei Ebenen ab, nämlich der Ebene von Textgrammatik und -semantik und der konzeptuellen Ebene: Auf der ersten Ebene tragen die syntaktische Position und Funktion des Ausdrucks (z. B. als Erstglied eines Satzes und als Subjekt), Agentivität (die semantische Rolle als Handelnder) und die Anzahl sowie die Art der Mehrfacherwähnung zur Erhaltung von Salienz bei. (Schwarz-Friesel & Consten 2014: 112)

Ein Referenz-Wechsel (**Switch-Reference**, Van Valin & LaPolla 1997: 287) geschieht einerseits bei Einführung eines neuen Referenten in Subjektposition, andererseits bei Re-Aktivierung eines im aktiven Teil des Gedächtnisses gegebenen Referenten.[25] In kontinuierlichen Phasen bleibt ein Referent über mehrere, in sprachlichen Äuße-

24 Topik hier als Aboutness-Topik (vgl. Matić 2015: 96); vgl. auch Skopeteas u. a. 2006: 2: „[…] we regard a 'topic' as a referent which the remainder of the sentence is about […] crucially followed by comment, typically containing a focus element. The topic has often been previously introduced into the discourse, but does not have to have been. We keep the notions of 'topic' and 'given' apart." Schulze 2018 bezeichnet dieses auf die Informationsstruktur des Satzes bezogene Topik als „Informational Topic" („Topic as a function of communicative interaction", Schulze 2018: 73).

25 Zum Zusammenhang von Re-Aktivierung und Aufmerksamkeitsstatus vgl. Schwarz-Friesel & Consten 2014: 111: „Im TWM kommt es bei einer Anapher immer zu einer Re-Aktivierung des bereits repräsentierten Referenten. In kognitiven Ansätzen wird der Aktivierungsstatus solcher Referenten auch als Salienz (Auffälligkeit) beschrieben (s. Chafe 1976, Givón 1983). Saliente Textreferenten sind diejenigen, die die höchste kognitive Aufmerksamkeit erhalten, also im TWM am stärksten aktiviert sind. Daher sind sie im TWM die wahrscheinlichsten Anknüpfungspunkte für anaphorische Beziehungen."

rungen kodierte Folgen von Ereignisvorstellungen im Zentrum der Aufmerksamkeit (des Bewusstseins); kognitiv wird hier in der aktuellen Textproduktion bzw. Textverarbeitung gewissermaßen die Perspektive des Referenten eingenommen (vgl. Dik 1991: 248, 260f.; 1997a: 64, 247ff.). Switch-Reference kann also als **Perspektivwechsel** verstanden werden (vgl. Schulze 2018: 85; 2019: 21, 28f.):

> Entsprechend des durch die Intrada festgelegten TWM [*von Volkserzählungen*] erfüllt der Text also die Erwartungen, wonach die Hauptprotagonisten im Vordergrund und damit im Zentrum der Aufmerksamkeit bleiben. Ein Wechsel der Perspektive kommt im Wesentlichen in der Konkurrenz zwischen den beiden Hauptakteuren vor. Grundsätzlich sind aber *switch reference*-Passagen relativ ‚kurzlebig' […]. (Schulze 2019: 29, Hervorhebung im Original)

Die verschiedenen Typen dieses informationsstrukturellen Textstruktur-Parameters der Perspektivierung sind in unterschiedlichen Operationalisierungen und Theorien beschrieben worden. Man kann für narrative Texte allgemein ausgehen von „Wechsel von Ko-Referenz und *Switch-Reference*" (Schulze 2020: 607, Hervorhebung im Original), also einer blockweisen **Topic-Continuity** (Givón 1983a; vgl. auch 3.3.2), d. h. längeren Phasen mit gleichbleibendem Referenten im Zentrum (dem dadurch als solchen mitdefinierten Hauptreferenten).

Im Rahmen der funktionalen Satzperspektive (Daneš 1970; Beneš 1973; vgl. Eroms 2000) kann man diese Art der Perspektivierung als Progression durch Blöcke mit durchlaufendem Thema auffassen (s. Daneš 1970: 76; vgl. Li & Thompson 1979: 313: *topic chaining*; Schwarz-Friesel & Consten 2014: 104ff.: thematische Kontinuität).[26]

Im sprachtypologischen Vergleich kann für bestimmte Sprachen ein morphologisches Switch-Reference-Markierungssystem festgestellt werden (mit *switch-subject*- und *same-subject*-Markern), das also den Wechsel des Referenten in Subjektposition overt kodiert (s. Van Valin & LaPolla 1997: 287f.). Mit Schulze 2020 wird in dieser Arbeit unabhängig vom Vorliegen einer solchen Markierung die Switch-Reference-Struktur eines Textes als informationsstruktureller Textstruktur-Parameter aufgefasst: Dieses Merkmal verbindet funktional-syntaktische Grundparameter (Subjekt als zentrales Argument) mit der Referenzsemantik, indem es den Änderungsverlauf der Referenten in der zentralen Position jedes Clauses bestimmt. Dieser Parameter wird hier entsprechend als kategoriale Sequenz von `SWITCH-` (*switch subject*) und `CONT`-*tags* (*same subject*) operationalisiert:

26 Für andere Genres wären als alternative TWM-Perspektivierungsschemata folgende Mustertypen im Sinne der Theorie der thematischen Progression der funktionalen Satzperspektive möglich: lineare Progression als sukzessiver Perspektivwechsel (hier wird die neue, nicht-topikale Information im nächsten Satz Aussagegegenstand, also Satz-Topik; s. Daneš 1970: 75f.) oder Progression mit abgeleitetem Thema (s. Daneš 1970: 76f.).

```
SWITCH-CONT-CONT-SWITCH-SPEECH-...
```

Auflistung 3.5: Beispiel für Switch-Reference-Verlauf (Abfolge der Subjekt-Referenzstatus)

3.8.3 Aufmerksamkeitsstruktur (Pragmatische Typik)

Dieser informationsstrukturelle Parameter bezieht sich auf die **Selektionsfunktion** der Aufmerksamkeitssteuerung (s. 1.1.1, Fußnote 3; Krummenacher 2020: 221f.), die bestimmt, welche der momentan in einer Äußerung gegebenen und dadurch im Bewusstsein aktivierten referentiellen oder relationalen Informationseinheiten die für den Aufbau des Text-Modells wichtige (d. h. neue oder kontrastive) Information ist.[27]

Im Gegensatz zu dem gedächtnisbezogenen Aktivierungsstatus neuer Topiks und ähnlich wie der Parameter der Perspektivierung bezieht sich dieser Parameter allgemein auf den Aufmerksamkeitsstatus von Informationseinheiten, nämlich, ob diese eine **neue Information** im Sinne einer durch Hinzufügung oder Änderung (bei kontrastivem Fokus) im Text-Modell zu aktualisierenden Information darstellen. Solche sprachlich über Fokussierungsmittel (s. u.) kodierten **Konstruktionsanweisungen** für die Anpassung des Text-Modells während der Textverarbeitung stellen einen zentralen Parameter der Informationsstrukturierung eines Text-Modells dar; vgl. dazu Schwarz-Friesel & Consten 2011: 352: „recipients create a mental model of the world described in a specific text and store it in episodic memory [...]. The general idea behind this is that verbal expressions serve as mental processing instructions for the recipient [...]."

Im Gegensatz zum Parameter der Perspektivierung, der den Änderungsstatus des Satz-Topiks kodiert – also sich z. B. auf Topic-Continuity-Strategien bezieht, d. h. darauf, ob ein Referent über mehrere Äußerungen hinweg als Aussagegegenstand im Zentrum der Aufmerksamkeit bleibt –, ist hier die Frage, wie die Aufmerksamkeit des Bewusstseins in der Verarbeitung einer sprachlichen Äußerung durch sprachliche **Fokussierungsmittel** (Fokusmarker, Wortstellung usw.) auf die rhematische, also für die Aktualisierung des kognitiven Text-Modells zentrale (insofern neue) Information über das Satz-Topik gelenkt wird (d. h. auf den **Comment**, vgl. Skopeteas u. a. 2006: 2).

Für Volkserzählungen ist hier prototypisch von einer Fokussierung der Hauptreferenten und ihrer Handlungen („Emphase markierter Situationen und ihrer Akteure",

27 Vgl. Krummenacher 2020: 221f. zum Fokus der Aufmerksamkeit: „Man spricht von der Spotlight-Metapher der A. [*Aufmerksamkeit*] [...]. Informationen innerhalb des A.fokus werden selektiert und können das Verhalten beeinflussen, Informationen außerhalb des Fokus werden dagegen ignoriert." Solche nicht-selektierten Informationen sind sog. Background-Informationen, s. folgender Abschnitt (vgl. Skopeteas u. a. 2006: 3).

Schulze 2020: 607) auszugehen, repräsentiert durch „Korrelation von fokalen Verfahren mit spezifischen Ereignissen/Akteuren“ (Schulze 2020: 607).[28] Als quantitativer TWM-Parameter kann für verschiedene **Fokussierungstypen** deren relative Textfrequenz bzgl. referentieller (oder auch relationaler) Einheiten berechnet werden (für Details zu den in dieser Arbeit berücksichtigten pragmatischen Rollen im obugrischen Korpus s. 5.3.5).

Fokussierungstypen:	FOC	CTR	REPEAT	MFOC	FRAME
relative Häufigkeit der Fokussierungstypen:	0.4	0.3	0.1	0	0.2

Tabelle 3.8: Beispiel einer Frequenzverteilung für Fokussierungstypen

3.8.4 Vordergrund-Hintergrund-Strukturierung

Ein informationsstruktureller TWM-Parameter bzgl. der relationalen Informationseinheiten (Ereignisvorstellungen) ist die Unterscheidung zwischen Vordergrund- und Hintergrund-Informationen (Schulze 2018: 213), vgl. dazu Ehrlich 1987: 363: „The foreground of a narrative text is defined as linguistic material which charts the progress of a narrative through time, while the background is durative and descriptive material which serves to embellish and elaborate upon the foreground […].“

Die Vordergrund-Ereignisse sind als „main events“ (Hopper 1979: 213) die zentralen, neuen Informationen über die Referenten im Register des Text-Modells und werden als solche in der Textverarbeitung, also beim Aufbau des kognitiven Text-Modells im Arbeitsgedächtnis, als dessen zentrale Relationen abgespeichert. Die Hintergrund-Ereignisse sind dagegen bekannte Informationen (Kontextualisierungen, bezugnehmend auf andere Wissensquellen wie Weltwissen, vgl. Schulze 2018: 213) oder sind zumindest – etwa bei Informationen, die die genauere Strukturierung des Bühnenhintergrunds des kognitiven Modells betreffen, vor dem die Haupthandlung stattfindet (vgl. Abschnitt 3.5, *cognitive map*) – nicht primär relevant für die Aktualisierung des textuell kodierten kognitiven Situationsmodells.[29]

Für Volkserzählungen kann man annehmen, dass diese prototypisch wenig Hintergrund-Information geben (vgl. Schulze 2019: 28), sich also auf die Vermittlung

28 Vgl. Schulze 2019: 29 bzgl. des informationsstrukturellen TWM-Prototyps von Volkserzählungen als geschlossenem Modell: „Ein wesentliches Moment des TWM von Volkserzählungen des Typs DT im Udischen ist in der Tat, soweit wie möglich auf gegebenes Wissen zurückzugreifen und möglichst wenige der außerhalb des entsprechenden TWM stehenden Informationen einzubetten. Dieses Moment wird auch darin erkennbar, dass der emphatische Fokus-Marker *-al*, der in der udischen Alltagssprache sehr präsent ist, nur vereinzelt und dann oft geblockt vorkommt […][.]“

29 Vgl. auch Hopper 1979: 213 (Hervorhebung im Original): „It is evidently a universal of narrative discourse that in any extended text an overt distinction is made between the language of the actual story line and the language of supportive material which does not itself narrate the main events. I refer to the former — the parts of the narrative which relate events belonging to the skeletal structure of the discourse — as FOREGROUND and the latter as BACKGROUND.“

der Haupthandlung konzentrieren. Sprachlich drückt sich die Differenzierung zwischen Vordergrund- und Hintergrund-Information (bzw. die Fokussierung wichtiger Vordergrundinformationen) z. B. über Tempuswechsel, Subordination oder Wortstellung aus.[30] So wurde in diskurstypologischen Studien (Hopper 1979; Schiffrin 1981; s. Myhill 2001: 168) das sog. **Temporal-Sequencing** (Labov 1972, als „narrative clause") als Parameter eingeführt (s. Myhill 2001: 168), also die Markierung von zentralen Ereignissen im narrativen Diskurs als Vordergrund-Information mit sprachlichen Mitteln wie dem historischen Präsens:

> According to this criterion, a clause is temporally sequenced if it has past time reference and refers to the next event in a story line (e. g. the second clause, but not the first, in *I was reading in the library and this guy came up to me ...*). The sequencing function has been related to alternations in word order, voice, and verb form. For example, Schiffrin (1981) shows that the English historical present is associated with temporally sequenced clauses, while Hopper (1979) shows that temporal sequencing is associated with the use of the verbal forms with a *di*-prefix in Malay. (Myhill 2001: 168, Hervorhebung im Original)

Es handelt sich hier demnach um eine pragmatische Funktion, Sätze des narrativen Diskurses mit Hilfe u. a. von TAM-Mitteln als *foreground*-Information zu markieren (TAM-**Foregrounding**, Myhill 2001: 168), also eine Distinktion von Haupt- und Hintergrundinformation zu treffen (s. auch Schulze 2019: 22f.). Vgl. dazu die Feststellung bei Schulze 2020: 614, „[…] dass Tempora in solchen Volkserzählungen keinen wirklichen Zeit-Aspekt ausdrücken. Besonders in orientalischen Sprachen dienen sie vor allem zur Trennung von Hintergrund-Informationen (Vergangenheit) und Vordergrund-Informationen (Präsens)." Temporal-Sequencing-Markierung ist also – neben den bereits oben angesprochenen Backgrounding-Strategien durch Subordination (s. 3.2.4; vgl. Dik 1997b: 431f.; Schulze 2018: 214f.; 2004a: 556) – eine weitere Strategie der sprachlichen Kodierung der Vordergrund-Hintergrund-Strukturierung von Texten, speziell in narrativen Diskursen.

30 Etwa Foregrounding über VS-Wortstellung im Deutschen, vgl. den Beginn des Märchens „Der Froschkönig oder der eiserne Heinrich" in der Sammlung der Brüder Grimm (2010: Bd. 1, Nr. 1, 29):
In den alten Zeiten, wo das Wünschen noch geholfen hat, (Background)
lebte (V) *ein König* (S) ... (Foreground).
Vgl. auch die Präsensmarkierung als Foreground-Marker in der obugrischen Erzählung 730, Sätze 4 und 18, hier in engl. Übersetzung mit Angabe der entsprechenden Sequencing-Markierung im obugrischen Originaltext:
In those days it was (nullmarkiert) *so cold* (Background)
that the ice in the lakes cracked (Präsens-markiert) *in winter.* (Foreground)

As I was looking around the marsh, I saw (nullmarkiert) (Background)
a man with reindeer riding at full speed (Präsens-markiert). (Foreground)

Ob eine Temporal-Sequencing-Strategie gegeben ist (für das Obugrische ist dies der Fall, s. Nikolaeva 1999: 26) und welcher Art diese ggf. ist, ist sprachspezifisch unterschiedlich. Allgemein kann man folgende Operationalisierung ansetzen:

$$\text{Temporal-Sequencing-Stärke (TEMP\_SEQ)} = \frac{\text{Anzahl Verbformen im Präsens}}{\text{Anzahl Verbformen}} \tag{3.25}$$

3.8.5 Textinterne Diskursstrukturen

Im Rahmen des Aufbaus kognitiver Text-Modelle in der Verarbeitung von Texten als Sequenzen von Sprachhandlungen haben Sequenzen direkter oder indirekter Rede einen hervorgehobenen informationsstrukturellen Status, da sie Situationsausschnitte textuell kodieren, die selbst aus Sprachhandlungen bestehen (s. Banfield 1973: 29) und deren Inhalt damit nicht direkt Teil der Haupthandlung ist, sondern ein eigenes, eingebettetes Text-Modell impliziert, das nicht notwendigerweise auf die Umstände der Situation bezogen sein muss: „Dialogic sequences […] do not contribute to the general dynamic segments in a text, but report about the verbal interaction of agonists within a given situation" (Schulze 2018: 203).

Während also die Vordergrund-Hintergrund-Strukturierung eine Unterscheidung der für den Situationsmodellaufbau relevanten Hauptinformationen gegenüber Zusatzinformationen für eine Einbettung und Anbindung des Modells in im Langzeitgedächtnis gespeicherte **Wissenskontexte** (Weltmodelle) betrifft, gehören **textinterne Diskursstrukturen** zwar zur Erzählebene, da diese textuell kodierten Sprachhandlungen Teil des Hauptgeschehens sind; sie nehmen aber im Aufbau des kognitiven Text-Modells eine Sonderrolle ein, da sie gewissermaßen einen eigenständigen **Text im Text** darstellen (vgl. Banfield 1973: 34) und entsprechend ein eigenes Text-Modell implizieren (*discourse layering*, s. Dik 1997b: 424ff.). Dieses eingebettete Text-Modell ist etwa bei narrativen Texten anderer Art als das Hauptmodell, nämlich ein Modell rhetorischer Interaktion (Schulze 2018: 190) innerhalb eines narrativen Text-Modells. Bei der Verarbeitung solcher eingebetteten Texte werden also ggf. weitere TWM aktiviert, insbesondere dialogbezogene (diese sind „grounded in frames of interactional typicity in a sociocultural milieu", Schulze 2018: 190). Das Auftreten von Dialogsequenzen in narrativen Texten kann also (als sprachlicher Ausdruck der Bezugnahme auf entsprechende andere TWM) als möglicher Parameter eines TWM dieses Texttypus gelten.

Damit ist die Verteilung von Abschnitten direkter oder indirekter Rede neben der Vordergrund-Hintergrund-Strukturierung ein weiterer mesostruktureller – also auf die mittlere Ebene zwischen Satz-Informationsstruktur und globaler Makrostruktur des Textes bezogener – Parameter, der mit der Einbettung von Sprachhandlungen in Texten eine hierarchische Gliederung im textuell (d. h. durch Sprachhandlungen) kodierten kognitiven Situationsmodell beschreibt. Es tritt hier also ein rekursives

Moment von Textstruktur auf (vgl. Dik 1997b: 427), wenn Sprachhandlungen Sprachhandlungen kodieren: „[…] there can be texts within texts, each set off from the containing text by the conventions for direct quotation“ (Banfield 1973: 34).

Die Frequenz und die sequentielle Verteilung von dialogischen Abschnitten kann z. B. über entsprechende Einleitungen durch Verben des Sagens operationalisiert werden (sog. *inquit*-Formeln). In dieser Arbeit wird der Parameter als Partitur-Folge über die im obugrischen Korpus annotierten Informationen zum diskursiven Status von Clauses berechnet, vgl. dazu auch Schulze 2020: 626: „Die Informationen zum Handlungsablauf werden wie für viele Märchen typisch in mehr oder minder starkem Umfang durch dialogische Szenen unterbrochen […][.] Die entsprechenden Partiturstimmen […] lassen entsprechende Szenen gut sichtbar werden […][.]“

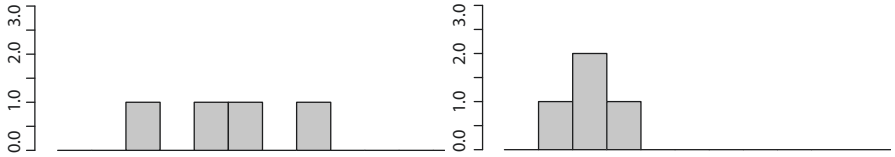


Abbildung 3.6: Binäre Kodierung des Rede-Status von Clauses vs. aggregierte Partitur-Kodierung der Anzahl von Clauses in direkter Rede pro Satz

SPEECH-Status von Prädikaten:	`NONE-NONE-SPEECH-NONE-SPEECH-SPEECH-NONE-SPEECH-NONE-...`
numerische Kodierung:	`0-0-1-0-1-1-0-1-0-0-0-0-0-0-0`
Partitur (Satz-aggregiert):	`0-1-2-1-0-0-0-0-0-0`

Tabelle 3.9: Beispiel für Kodierungen textinterner Diskursstrukturfolgen

Es ist zu erwarten, dass die Art (direkte vs. indirekte Rede),[31] die Stärke, die sequentielle Verteilung sowie ggf. auch die Rekursionstiefe[32] von solchen **textinternen Diskursstrukturen** mit dem Textgenre variiert. So zeichnen sich speziell narrative Texte typischerweise aus durch über den Text verteilte Dialog-Passagen, die auf die in der Haupthandlung stattfindenden Ereignisse und Orte Bezug nehmen, diese also reflektieren, aber nicht selbst dazu beitragen (vgl. Schulze 2018: 203, 212; 2020: 607, 626): „[…] these passages contribute only little to the action chain itself, both rather anticipates certain actions or comment upon them“ (Schulze 2018: 212).

31 In Kapitel 6 wird nur die direkte Rede berücksichtigt, da im obugrischen Korpus nur dieser Typ ausgezeichnet ist (und dies auch die vorherrschende Art der Rede in den Texten ist).

32 Vgl. etwa die Verwendung textinterner Diskursstrukturierung als stilistisches literarisches Mittel in den Romanen von Thomas Bernhard (Staffelung von Ebenen indirekter Rede).

4 Methoden einer quantitativen Texttypologie

Kapitelzusammenfassung

In diesem Kapitel werden Methoden der quantitativen Linguistik sowie der Computerlinguistik für eine automatische Genre-Klassifizierung von Texten anhand von kognitiv begründeten, textstrukturellen Parametern vorgestellt. Dazu gehören Repräsentationsmethoden wie das Datenrepräsentationsmodell, das die Formalisierung textstruktureller Parameter über Feature-Sets in einem n-dimensionalen Merkmalsvektorraum ermöglicht, sowie sequentielle Textstruktur-Repräsentationen für die Operationalisierung von globalen und lokalen kategorialen Sequenzmustern sowie von numerischen Partitur-Folgen. Es werden Verfahren der Feature-Construction und der Feature-Extraction aus dem Bereich des maschinellen Lernens und des Data-Mining eingeführt, die zur Erzeugung solcher Repräsentationen als Operationalisierungen der in Kapitel 3 diskutierten TWM-Parameter geeignet sind. Ebenso werden Klassifizierungsmethoden besprochen, die eine Typisierung und Feature-Analyse dieser textstrukturellen Repräsentationen ermöglichen: Die hierarchische Clusteranalyse erlaubt über die Entdeckung von geteilten Strukturmustern in den textstrukturellen Repräsentationen eine induktive Typisierung von Texten gemäß der TWM-Parameter (Feature-Exploration). Mit Klassifikationsmethoden wie dem Random-Forest-Klassifikator kann die Fähigkeit von TWM-Parametern zur Differenzierung verschiedener Genrekategorisierungen analysiert werden (Feature-Selection). Die Darstellung orientiert sich dabei an dem in Abschnitt 1.2 formulierten Anspruch, die relevanten mathematisch-formalen Aspekte so aufzubereiten, dass die mathematischen Grundlagen statistischer Lernmethoden für Mustererkennung, Feature-Exploration und Feature-Analyse einer sprachwissenschaftlichen Leserschaft erfolgreich vermittelt werden.

4.1 Datenrepräsentationsmodell

4.1.1 Feature-Set-Repräsentation

In der quantitativen Linguistik und der Computerlinguistik wird für die Repräsentation von Texten in Anwendungen wie dem Dokumentenclustering oder der Textklassifikation üblicherweise das sog. **Datenrepräsentationsmodell** verwendet (s. Manning & Schütze 1999: 495, 576; Mehler 2005: 341). Dabei werden die Texte als die zu klassifizierenden Objekte anhand spezifischer numerischer Merkmale über ein **Feature-Set** repräsentiert, d. h. in Form einer Menge von Merkmal-Wert-Paaren, die als Datensätze (Zeilen) einer Datenmatrix zusammengefasst werden (s. Tabelle 4.1 für ein Beispiel).

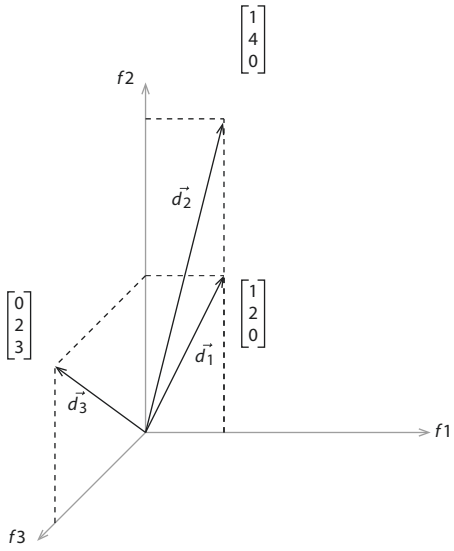


Abbildung 4.1: Dreidimensionaler Feature-Space: Vektoren als Objektrepräsentationen

	Merkmal A (f_1)	**Merkmal B** (f_2)	**Merkmal C** (f_3)
Text 1 ($\vec{d_1}$)	1	2	0
Text 2 ($\vec{d_2}$)	1	4	0
Text 3 ($\vec{d_3}$)	0	2	3

Tabelle 4.1: Feature-Set-basierte Text-Repräsentationen

Mathematisch können solche Feature-basierten Objektrepräsentationen des Datenrepräsentationsmodells als Vektoren in einem durch die Merkmale als Dimensionen aufgespannten Raum verstanden werden (s. Abbildung 4.1). Die Dimensionalität dieses Merkmalsraums (**Feature-Space**) entspricht also der Anzahl der in einem solchen Feature-Set zusammengefassten Merkmale (und kann je nach zugrunde liegender Operationalisierung hochdimensional sein). Man spricht hier auch von einem durch die Merkmalsdimensionen gebildeten **Vektorraum**, in dem Objekte über n-dimensionale numerische Vektoren, sog. Merkmalsvektoren, repräsentiert sind (Vektorraummodell, vgl. Manning & Raghavan & Schütze 2009: 120). Die Skala einer Dimension eines solchen Merkmalsraumes entspricht der Skala des diese Dimension durch seine Werteverteilung aufspannenden Merkmals.

Das theoretische Konstrukt eines Merkmalsraums erlaubt durch die mathematische Formalisierung einer Menge multivariater Variablen als Feature-Set die Anwendung von Klassifizierungsmethoden, insbesondere **Clustering** zum automatischen Auffinden von Gruppen ähnlicher Objekte über die Berechnung von Distanzen zwischen deren Merkmalsvektorrepräsentationen im Feature-Space sowie **Klassifikation** als Lernen einer Abbildung von Feature-Vektoren auf Klassenlabels (vgl. Manning & Raghavan & Schütze 2009: 120).

Von den in in dieser Arbeit zur textstrukturellen Repräsentation als TWM-Parameter verwendeten **globalen** textstrukturellen Parametern (Abschnitte 3.2–3.4), die textweite Maßzahlen sind, können beliebige Kombinationen dieser Variablen als Merkmale in einem Feature-Set zusammengefasst werden. Da hier voneinander unabhängige, globale Maße, die textstrukturelle Eigenschaften in einer Kennzahl zusammenfassen (z. B. Type-Token-Verhältnis, Lexikalische Dichte), zu einem Modell kombiniert werden (als Dimensionen eines Feature-Raums angesetzt werden), ist bei diesen globalen Feature-Sets eine Skalierung notwendig (s. 4.1.3.2), da die einzelnen

Parameter – im Gegensatz zu den folgenden *bag*-Modellen – keine gemeinsame Skala besitzen.

Report 4.1.1[1] zeigt ein Beispiel für die Repräsentation von Texten des obugrischen Korpus über ein Feature-Set solcher globaler, textstrukturell-quantitativer Parameter (in dieser Tabelle sind die Feature-Werte noch nicht skaliert, siehe Report 4.1.3 im Abschnitt zur Standardisierung für die skalierte Version des Feature-Sets).

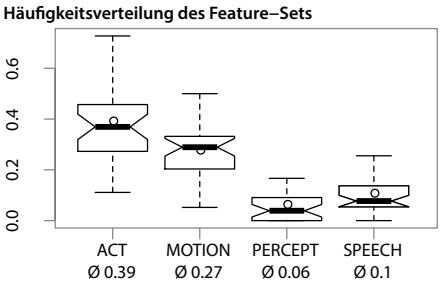
Text-ID	CL_ELAB	CL_COMPLEX	SENT_COMPLEX	RED	LEX_DENS
728	3.27	0.15	0.31	1.45	6.37
730	3.04	0.18	0.33	1.54	7.79
732	2.94	0.19	0.33	1.72	7.81
741	2.25	0.00	0.00	1.10	3.02
742	2.63	0.08	0.18	1.46	8.21
750	2.50	0.08	0.19	2.34	8.54

Report 4.1.1: Unskaliertes Feature-Set globaler Parameter (Beispiel)

Dagegen sind die **textfunktional** motivierten TWM-Parameter, die sich auf die von Informationseinheiten gebildete Struktur des kognitiven Text-Modells beziehen (Abschnitte 3.6–3.8), typischerweise keine textweiten Merkmale, sondern messen die Frequenzen der Ausprägungen bestimmter funktionaler Eigenschaften eines Informationseinheitstyps, so etwa die relativen Textfrequenzen verschiedener Verbklassen als funktionale Merkmale der Ereignistypik eines Textes (vgl. Report 4.1.2). Konkret bilden hier also die Labels (*tags*) von entsprechenden funktionalen Annotationsgrößen die Dimensionen des Feature-Sets.[2]

Text-ID	ACT	MOTION	PERCEPT	SPEECH
728	0.37	0.40	0.11	0.00
730	0.33	0.32	0.03	0.03
732	0.37	0.30	0.03	0.10
741	0.25	0.75	0.00	0.00
742	0.37	0.27	0.06	0.14
750	0.23	0.34	0.03	0.26

Report 4.1.2: Bag-of-Tags-Feature-Set von Ereignistypen (Beispiel)



Plot 4.1.1: Boxplot des Feature-Sets von Ereignistypen (Beispiel)

1 Die Report-Tabellen in dieser Arbeit geben üblicherweise nur die ersten Einträge (Zeilen) des jeweiligen Datensatzes an; bei Feature-Sets wie in Report 4.1.1 sind dies die Feature-Werte der ersten Texte im entsprechenden, nach Text-ID geordneten Datensatz.

2 Für eine zusammenfassende Statistik können Feature-Sets u. a. durch sog. Boxplots als graphische Darstellung deskriptiver Lage- und Streuungsmaße einer Verteilung visualisiert werden, wie in Plot 4.1.1 für das Ereignistypik-Feature-Set; die Begrenzungen der Box entsprechen dabei dem oberen bzw. unteren Quartil, der Strich in der Box dem Median und der Punkt dem arithmetischen Mittel.

Solche von der linearen Anordnung abstrahierenden, frequenzbezogenen Objektrepräsentationen stellen formal eine **Multimenge** dar (auch: ***bag***, s. Manning & Schütze 1999: 495, 575f.; Deza & Deza 2016: 58). Eine Multimenge ist eine ungeordnete Elementsammlung, die – im Gegensatz zu einer Menge – dasselbe Element mehrfach enthalten kann, beispielsweise zwei Vorkommen der Verbklasse `ACT` in folgender Multimenge: `(ACT, ACT, MOTION)`. Mathematisch kann man unter einer Multimenge auch eine Abbildung der verschiedenen Elemente (der enthaltenen Types) auf die Frequenz ihres Vorkommens in der Multimenge verstehen, im Beispiel: `(ACT:2, MOTION:1)`. Als Werttypen von *bag*-Feature-Sets können – wie beispielsweise im Ereignistypik-Feature-Set in Report 4.1.2 – statt absoluter Frequenzen auch relative Häufigkeiten verwendet werden. Ebenso ist auch eine Gewichtung der Frequenzdaten zum Zwecke einer Normalisierung möglich. Die Grundidee des *bag*-Modells bleibt aber auch hier bestehen: Es sind quantitative Modelle mit den Types als Merkmalen, die also von der linearen Anordnung abstrahieren.

Bag-of-Words-Modelle (s. Manning & Raghavan & Schütze 2009: 117; Beyerer & Richter & Nagel 2017: 89) sind klassische Repräsentationsformen für Dokumente in der Textklassifikation des Information-Retrieval, das Dokumente nach relevanten Suchtermen klassifiziert. Hier werden Texte also über die Frequenzen ihrer Wortterme repräsentiert (Aggarwal 2018: 2). Diese Frequenzdaten werden dabei üblicherweise bzgl. der Termrelevanz gewichtet (Tf-idf-Maß, s. Deza & Deza 2016: 341; Manning & Raghavan & Schütze 2009: 117f.). Analog zu diesen Bag-of-Words-Feature-Sets, deren Feature-Dimensionen durch das Wort-Vokabular der Dokumentensammlung gebildet werden, werden die in dieser Arbeit relevanten **textstrukturellen** *bag*-Feature-Sets als **Bag-of-Tags** bezeichnet, da ihre Feature-Dimensionen durch das *tag*-Vokabular der funktionalen Annotationskategorien des jeweiligen textstrukturellen Parameterbereichs gebildet werden (referenz-, relationssemantische oder informationsstrukturelle *tags*; vgl. 3.1.1).[3]

4.1.2 Feature-Construction und -Extraction

4.1.2.1 Feature-Construction-Methoden

Der Begriff der **Feature-Construction** (s. Motoda & Liu 2002: 69) bezieht sich typischerweise auf Methoden der Berechnung neuer Merkmale auf einem vorhandenen Feature-Set mit dem Ziel von dessen Optimierung:

> Feature construction is a process that discovers missing information about the relationships between features and augments the space of features by inferring or creating addi-

3 Vgl. auch Beyerer & Richter & Nagel 2017: 88ff. bzgl. der Adaption des Bag-of-Words-Modells der NLP für die Mustererkennung in anderen Bereichen.

> tional features [...]. Assuming there are n features $A_1, A_2, ..., A_n$, after feature construction, we may have additional m features $A_{n+1}, A_{n+2}, ..., A_{n+m}$. (Motoda & Liu 2002: 69)

In dieser Arbeit wird Feature-Construction als allgemeiner Begriff für die Konstruktion von neuen Feature-Variablen durch Anwendung von arithmetischen Rechenoperationen (**konstruktive Operatoren**, s. Motoda & Liu 2002: 69; Beyerer & Richter & Nagel 2017: 39ff.) auf gegebene basale Merkmale verstanden. Die für die Fallstudie einer quantitativen kognitiven Texttypologie notwendigen Feature-Construction-Berechnungen sind bereits in Kapitel 3 behandelt und werden in Kapitel 6 nochmals im Kontext der Auswertungen vorgestellt.

4.1.2.2 Feature-Extraction-Methoden

Als **Feature-Extraction** (s. Motoda & Liu 2002: 68f.) wird der Vorgang bezeichnet, ein gegebenes Feature-Set durch **Transformations- und Aggregationsmethoden** in ein neues, kompakteres Feature-Set zu überführen (Guyon u. a. 2006: IX; vgl. auch Liu & Motoda 1998: 4):

> Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [...]. Assuming there are n features or attributes $A_1, A_2, ..., A_n$, after feature extraction we have another set of new features $B_1, B_2, ..., B_m (m < n), B_i = F_i(A_1, A_2, ..., A_n)$ and F_i is a mapping function. [...]. The goal of feature extraction is to search for a minimum set of new features via some transformation according to some performance measure. (Motoda & Liu 2002: 68f.)

Ein durch Feature-Extraction-Methoden dimensionsreduziertes Feature-Set ist einerseits effizienter, z. B. in der Verwendung in Klassifizierungsanwendungen (s. Guyon u. a. 2006: IX) – außerdem kann die Feature-Extraction eingesetzt werden, um für eine Forschungsfrage relevante Parameter zu gewinnen, was in dieser Arbeit für die Evaluation der TWM-Parameter wichtig wird. Entsprechend wird hier unter dem Begriff der Feature-Extraction die Aggregation von textstrukturellen Parametern aus Merkmalswerten verstanden, die zuvor aus den basalen Annotationsdaten im Feature-Construction-Prozess berechnet wurden. Die Feature-Extraction dient dabei der Generierung von Feature-Sets, die für eine TWM-Feature-Analyse geeignet sind. Als Beispiel kann die Aggregation von Daten zur referentiellen Distanz für einzelne Referentenvorkommen durch Mittelwertbildung bzgl. der jeweiligen grammatischen Relation gelten (s. 6.3.1).

Zu den **Transformationsmethoden** zählt u. a. die Transformation des Skalenniveaus von Merkmalen, also z. B. eine Diskretisierung von kontinuierlichen Werten (s. Tan & Steinbach & Kumar 2006: 57ff.; Aggarwal 2015: 475). **Aggregationsmethoden** dienen der Vereinigung von Merkmalen in einem gebündelten Merkmal (vgl. Tan & Steinbach & Kumar 2006: 45) über Methoden wie etwa Durchschnitts-

bildung über ein numerisches Merkmal bzgl. der Niveaus eines zweiten, kategorialen Merkmals.

Anschließend erfolgt die Zusammenstellung der zuvor konstruierten und extrahierten Feature-Vektoren der Länge n in einem **Feature-Set**. Dazu werden die einzelnen n-dimensionalen Textrepräsentationen zu einer m x n-Datenmatrix mit m Texten (Zeilen) und n Merkmalen (Spalten) kombiniert, wobei jeder Zeilenvektor die Merkmalsvektorrepräsentation eines Textes im n-dimensionalen Feature-Space darstellt.[4] Diese Feature-Set-Datenmatrix kann dann als Input für Feature-basierte Klassifizierungsmethoden dienen.

Generell kann bei der Konstruktion und Extraktion von Features unterschieden werden zwischen *data-driven*-Ansätzen, in denen neue Features durch mathematische Methoden aus den vorhandenen Daten gewonnen werden, sowie *hypothesis-driven*- bzw. *knowledge-based*-Ansätzen, in denen neue Features basierend auf Hypothesen bzw. *domain knowledge* erzeugt werden (Motoda & Liu 2002: 70). Die Feature-Construction in dieser Arbeit folgt letzteren Ansätzen, da die hier aufzubauenden Analyse-Feature-Sets kognitiver Parameter eine Operationalisierung dieser Parameter, basierend auf der Grundhypothese zur kognitiven Texttypologie (Text-Weltmodelle, s. Abschnitt 1.2), anstreben. Die Konstruktion und Extraktion der verschiedenen Textstrukturmuster aus den Korpusdaten geschieht dementsprechend also hypothesenbasiert:

> Auch wenn eine korpuslinguistische Diskursanalyse induktiv vorgeht, vorurteilslos tut sie es nicht. Wenn der Ausgangspunkt rekurrente lexikalische Elemente sind, die in die Beschreibung von Sprachgebrauchsmuster münden, wird mit diesem Ausgangspunkt bereits eine starke Hypothese vorausgesetzt, die solche Sprachgebrauchsmuster als grundlegende Indikatoren für Diskurse annimmt. (Bubenhofer 2008: 432)

Die Konstruktion der zu untersuchenden Merkmale in dieser Arbeit basiert also auf spezifischen Grundannahmen zur Textverarbeitung durch die menschliche Kognition – die darauf aufbauende Untersuchung zur datengestützten Rekonstruktion von TWM-Modellen über Clusteranalysen geht aber als explorative Mustersuche grundsätzlich induktiv vor (*corpus-driven*, vgl. Bubenhofer 2008: 411f.), sie sucht also – basierend auf der theoretischen Prämisse eines analogen Aufbaus dieser zu untersuchenden kognitiven Modelle als *usage-based-models* (Langacker 2000) über Induktion aus Sprachgebrauchsdaten durch die Kognition – in den textstrukturellen Daten nach Strukturmustern und setzt somit keinen Mustertyp als Hypothese an (hypo-

4 Dabei können zwei Texte theoretisch dieselbe Feature-Set-Repräsentation haben, d. h. die in der Datenmatrix gespeicherten Feature-Set-Repräsentationen der zu untersuchenden Menge von Texten bilden wiederum eine Multimenge (s. Manning & Schütze 1999: 495).

thesenprüfend), sondern geht stattdessen **hypothesenbildend** vor, vgl. noch einmal Bubenhofer:

> Korpuslinguistische Methoden ersetzen nicht bestehende diskurslinguistische Methoden, sondern ergänzen sie. Allerdings setzen sie an einem grundlegenden Punkt von Diskurslinguistik ein: Eine corpus-driven operierende Korpuslinguistik geht induktiv, und damit hypothesenbildend, vor. Statt nur als Hilfsmittel zur Hypothesenüberprüfung zu dienen, verhilft sie der Diskursanalyse zu einem anderen Startpunkt, in dem zunächst ein Korpus auf seinen musterhaften Sprachgebrauch untersucht wird. (Bubenhofer 2008: 431)

4.1.3 Normalisierung und Standardisierung

4.1.3.1 Normalisierung

Normalisierung und Skalierung sind wichtige Schritte in der Feature-Extraction; ihr Zweck ist es, die Objektrepräsentationen vergleichbar zu machen (s. Beyerer & Richter & Nagel 2017: 32). Bei der **Normalisierung** geht es dabei um die Herstellung einer Vergleichbarkeit von Objekten (x) unterschiedlicher Größe ($\|x\|$) in einer spezifischen Feature-Dimension, in der diese Objektgrößen keinen Einfluss haben sollen:

$$x' = \frac{x}{\|x\|} \tag{4.1}$$

Vgl. dazu Guyon & Elisseeff 2006 mit einem Beispiel aus der Bildverarbeitung:

> Consider for example the case where x is an image and the x_i's are the number of pixels with color i, it makes sense to normalize x by dividing it by the total number of counts in order to encode the distribution and remove the dependence on the size of the image. This translates into the formula: $x' = x/\|x\|$. (Guyon & Elisseeff 2006: 3)

In der Klassifizierung von Textobjekten wird üblicherweise eine Normalisierung bzgl. der Textlänge vorgenommen. Entsprechend verwendet diese Arbeit textlängenbezogene Größen wie die relative Textfrequenz textueller Informationseinheiten oder Text-Durchschnittswerte bestimmter textstruktureller Features, um aus unterschiedlichen Textlängen resultierende Verzerrungen auszugleichen. Diese Textlängennormierung geschieht dabei – für jedes Feature unabhängig – schon im Feature-Construction-Prozess, also vor einer etwaigen Standardisierung des Feature-Sets (vgl. Guyon & Elisseeff 2006: 3; s. auch Beyerer & Richter & Nagel 2017: 32).

Bei Sequenzen geschieht entweder eine inhärente Normierung unterschiedlicher Sequenzlängen während der Berechnung der Distanzmatrix (s. 4.1.4) durch Methoden wie das Dynamic-Time-Warping (Beyerer & Richter & Nagel 2017: 38f.; s. auch 4.4.2.2), oder es sind, wie bei Optimal-Matching-Distanzen für kategoriale Sequenzen, verschiedene Normalisierungen der Abstände zwischen zwei Sequenzen über

die Sequenzlänge möglich (etwa *maxlength* als Normierung durch die Länge der längeren Sequenz, s. Gabadinho u. a. 2011: 29; s. auch 4.4.2.1). Häufige Teilfolgen (Frequent-Patterns, s. 4.4.4) können als frequenzbezogene Merkmale in einem Feature-Set verwendet und entsprechende Feature-bezogene Normalisierungen angewendet werden; in Abschnitt 6.4.2 wird für die Normalisierung häufiger Ereignisabfolgen statt einer Normierung, z. B. über die Anzahl von Zuständen in einem Text, eine binäre Operationalisierung bzgl. des Vorhandenseins eines Musters (Presence/Absence) in einem Text gewählt, die eine solche Längennormierung überflüssig macht.

4.1.3.2 Standardisierung

Die **Standardisierung** eines Feature-Sets erreicht durch Zentrierung sowie Skalierung der Dimensionen des Merkmalsraums die Vergleichbarkeit unterschiedlich skalierter, in einem Feature-Set zusammengefasster Merkmale:

> Features can have different scales although they refer to comparable objects. Consider for instance, a pattern $x = [x_1, x_2]$ where x_1 is a width measured in meters and x_2 is a height measured in centimeters. Both can be compared, added or subtracted but it would be unreasonable to do it before appropriate normalization. The following classical centering and scaling of the data is often used: $x'_i = (x_i - \mu_i)/\sigma_i$, where μ_i and σ_i are the mean and the standard deviation of feature x_i over training examples. (Guyon & Elisseeff 2006: 3)

Eine häufig verwendete Metrik für die Anpassung der Skalen der Merkmalsvariablen eines Feature-Sets auf ein einheitliches Niveau ist die sog. **z-Standardisierung** (vgl. Guyon & Elisseeff 2006: 3):

$$x' = (x - \mu)/\sigma \tag{4.2}$$

In der z-Standardisierung wird eine Zentrierung der Verteilung der Werte einer Feature-Dimension auf deren Mittelwert μ vorgenommen, indem die Werte durch Subtraktion um diesen Mittelwert verschoben werden, sodass der neue Mittelwert der angepassten Verteilung bei 0 liegt. Zusätzlich erfolgt eine Skalierung der Werte dieser Dimension gemäß deren Streuung durch Division durch die Standardabweichung $\sigma = \sqrt{\sum_{j=1}^{n}(x_j - \mu)^2}$.

	Feat. A	Feat. A skal.	Feat. B	Feat. B skal.
Text 1	1	−1	10	−1
Text 2	2	0	20	0
Text 3	3	1	30	1
Mittelwert μ:	2		20	
Standardabw. σ:	1		10	

Tabelle 4.2: Beispiel z-Standardisierung anhand zweier Feature-Dimensionen

Während also die Normalisierung für jedes Objekt (Zeile eines Feature-Sets) eine objektspezifische Gewichtung seiner Feature-Werte vornimmt (abhängig von der ‚Größe' der Objekte, also hier von der Textlänge), vereinheitlicht die Standardisierung die Werte aller Objekte pro Feature-Dimension (d. h. der Spalten eines Feature-Sets) basierend auf deren Verteilung (Mittelwert und Streuung), sodass die skalierten Feature-Dimensionen eine vergleichbare Skala besitzen und somit ein skalennormierter, **homogener Feature-Space** entsteht (vgl. Beyerer & Richter & Nagel 2017: 18f.).

Die Skalierung wird entsprechend *nach* Erstellung und Längennormierung des Feature-Sets vorgenommen; notwendig ist sie in Fällen wie den globalen Feature-Sets, die Merkmale mit unterschiedlichen Skalen besitzen, also einen heterogenen Feature-Raum mit nicht vergleichbaren Skalen aufweisen (vgl. Beyerer & Richter & Nagel 2017: 18f.). Allgemein hängt der Einsatz einer Skalierung also sowohl von den Daten als auch von der Fragestellung ab (vgl. James u. a. 2017: 398). Report 4.1.3 zeigt das Ergebnis der z-Standardisierung des globalen Feature-Sets von oben (Report 4.1.1).

Text-ID	CL_ELAB	CL_COMPLEX	SENT_COMPLEX	RED	LEX_DENS
728	0.33	−0.04	0.43	−0.58	0.06
730	−0.03	0.19	0.55	−0.34	0.83
732	−0.18	0.27	0.55	0.11	0.84
741	−1.23	−1.30	−1.44	−1.50	−1.77
742	−0.65	−0.65	−0.35	−0.56	1.06
750	−0.85	−0.65	−0.32	1.75	1.24

Report 4.1.3: Skaliertes Feature-Set globaler Parameter (Beispiel)

4.1.4 Distanzmaße für Feature-Sets

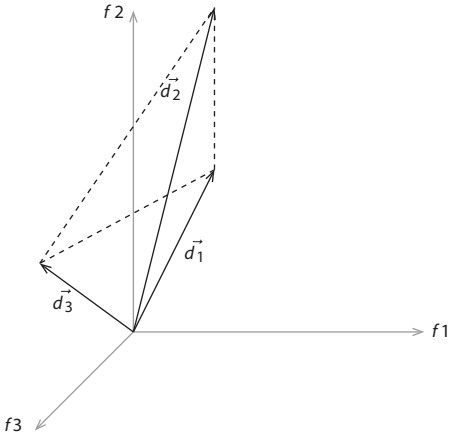


Abbildung 4.2: Euklidischer Abstand zwischen Vektoren in dreidimensionalem Raum

	Text 1 ($\vec{d_1}$)	**Text 2** ($\vec{d_2}$)	**Text 3** ($\vec{d_3}$)
Text 1 ($\vec{d_1}$)	0		
Text 2 ($\vec{d_2}$)	2.0	0	
Text 3 ($\vec{d_3}$)	3.16	3.74	0

Tabelle 4.3: Distanzmatrix mit euklidischem Abstand

Die zentrale Idee des Vektorraummodells besteht in der Interpretation der räumlichen Nähe zwischen den über Feature-Vektoren als Datenpunkte im n-dimensionalen Feature-Space repräsentierten Objekten als **Ähnlichkeit** bzgl. der Parameter des Vektorraummodells (vgl. Manning & Schütze 1999: 503, 540; Aggarwal 2015: 63ff.). Entsprechend wird in der automatischen Klassifizierung die **Distanz** zwischen Objektrepräsentationen eines Feature-Sets als ein invertiertes Ähnlichkeitsmaß (*dissimilarity measure*) berechnet. Die systematische Berechnung aller paarweisen Distanzen zwischen den m Feature-Vektoren eines Feature-Sets (m Zeilen = Datensätze) ergibt eine m×m-**Distanzmatrix** (auch: *dissimilarity matrix*; vgl. Aggarwal 2015: 169f.). Tabelle 4.3 zeigt die Distanzmatrix zu obigem Feature-Set in Tabelle 4.1, berechnet mit dem euklidischen Distanzmaß (s. 4.1.4.1).

4.1.4.1 Euklidisches Distanzmaß

Zur Berechnung der Abstände im Feature-Space können unterschiedliche Distanzmetriken Anwendung finden. Das einfachste Distanzmaß ist der **euklidische Abstand**, der die Distanz zwischen zwei durch Vektoren $\vec{x}$, $\vec{y}$ definierten Punkten in einem (n-dimensionalen) Raum als die Länge des Differenzvektors $\vec{z}$ zwischen den zwei Vektoren definiert, die sich über die sog. L2-Norm (auch: euklidische Norm) als $\|\vec{z}\|_2 = \sqrt{\sum_{i=1}^{n}(z_i)^2}$ berechnet (s. Deza & Deza 2016: 103):

$$dist(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\|_2 = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{4.3}$$

Das euklidische Distanzmaß fungiert in dieser Arbeit als Standardmaß für die Berechnung von Abständen zwischen Feature-basierten textstrukturellen Objektrepräsenta-

tionen, ggf. unter Verwendung von Normierungs- und Standardisierungsverfahren. Ein beim euklidischen Abstand als Distanzmaß auftretendes Problem ist seine Längensensibilität; in Abbildung 4.4 etwa haben die kurzen Vektoren entgegengesetzte Richtungen, aber ihr Punktabstand ist viel kleiner als der der beiden Vektoren mit ähnlicher Richtung. Neben der Möglichkeit der Verwendung der Kosinusähnlichkeit (s. Deza & Deza 2016: 341) als Distanzmaß – die in diesem Fall ohne Normierung eingesetzt werden kann, da dieses Ähnlichkeitsmaß nur die Richtung zweier Vektoren vergleicht (den eingeschlossenen Winkel), nicht die Länge der Vektoren – kann das Problem der Längensensibilität des euklidischen Distanzmaßes durch **Normalisierung** behoben werden (z. B. bzgl. unterschiedlicher Textlängen; s. 4.1.3.1; vgl. Beyerer & Richter & Nagel 2017: 32).

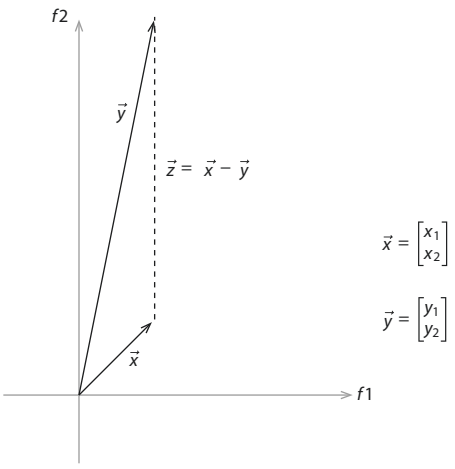


Abbildung 4.3: Beispiel zur Berechnung des euklidischen Abstands zwischen zwei Vektoren

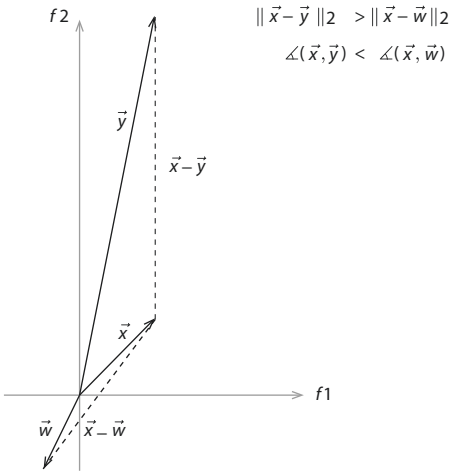


Abbildung 4.4: Beispiel für Längensensibilität

Ebenso kann der Einsatz des euklidischen Distanzmaßes beim Clustering sog. heterogener Feature-Räume (vgl. Beyerer & Richter & Nagel 2017: 18f.) zu einer Verzerrung führen, da hier – wie in den globalen textstrukturellen Feature-Sets in dieser Arbeit – Merkmale mit disparaten Skalen kombiniert sind; dieses Problem kann aber, wie in 4.1.3.2 angesprochen, durch **Standardisierung** beseitigt werden (also durch eine Skalierung und Zentrierung der Dimensionen des Merkmalsraums).

Zusammenfassend müssen in dieser Arbeit bei der Verwendung des euklidischen Distanzmaßes also globale Feature-Sets, die einen heterogenen Merkmalsraum aufweisen, nicht nur längennormiert, sondern auch skaliert werden, um die Merkmale vergleichbar zu machen. Dagegen spannen längennormalisierte *bag*-Feature-Sets

einen homogenen Feature-Space auf, hier kann somit auf eine Skalierung verzichtet werden.

4.1.4.2 Distanzmaße für sequentielle Daten

Während das euklidische Distanzmaß definiert ist als Abstandsmaß zwischen Vektoren in n-dimensionalen Merkmalsräumen, sind zur Berechnung des Abstands zwischen sequentiellen Textrepräsentationen, also kategorialen oder numerischen Wertfolgen, andere Typen von Distanzmaßen notwendig: Das sind einmal auf kategorialen Folgen definierte Editierdistanzmaße (**Edit-Distance**), die den Abstand zwischen zwei Zeichenfolgen als den Aufwand der notwendigen Änderungsschritte zur Transformation der einen Sequenz in die andere bestimmen (s. 4.4.2.1), sowie auf numerischen Folgen definierte Abstände wie die **Dynamic-Time-Warping**-Distanz, die den Aufwand der Abbildung einer Sequenz auf eine andere berechnet (s. 4.4.2.2).

4.2 Clusteringmethoden

4.2.1 Clustering als explorative Klassifizierung

Als Verfahren der explorativen Statistik teilt die **Clusteranalyse** Objekte *datenbasiert*, also auf Grundlage ihrer Merkmalsausprägungen, in Gruppen mit ähnlichen Merkmalsausprägungen ein (s. Jain & Murty & Flynn 1999; Tan & Steinbach & Kumar 2006: 487ff.; Aggarwal 2015: 205ff.; James u. a. 2017: 373ff.). Da hier induktiv eine automatische Klassifizierung von Objektrepräsentationen ohne vorher bekannte Klasseneinteilung durchgeführt wird, spricht man auch von unüberwachter Klassifizierung (***unsupervised***; vgl. Manning & Schütze 1999: 495ff.). Durch die in einer Clusteranalyse stattfindende Exploration der Lage der Datenpunkte im Feature-Space können unbekannte Strukturmuster in den Daten aufgedeckt werden (sog. **Feature-Exploration**) und so induktive Clustertypologien im Sinne von Gruppen ähnlicher (im Merkmalsraum benachbarter) Punkte berechnet werden.

Als die zwei Haupttypen von Clusteringalgorithmen sind die hierarchische und die partitionierende Clusteranalyse zu unterscheiden; während die **partitionierende** Clusteranalyse von einer festen Partitionierung der Datenpunkte im Merkmalsraum ausgeht (vorgegebene Anzahl an Clustern) und diese Partitionierung iterativ optimiert (s. Jain & Murty & Flynn 1999: 278ff.), geht die **hierarchische** Clusteranalyse von der feinsten oder gröbsten Partition aus und berechnet schrittweise Clustergruppen (Mengen von ähnlichen Datenpunkten), basierend auf den Distanzen der Clusterelemente (s. Manning & Schütze 1999: 500f.; Jain & Murty & Flynn 1999: 267, 275ff.).

Das hierarchische Clustering hat gegenüber der partitionierenden Clusteranalyse u. a. den Vorteil, dass keine Vorgabe der Anzahl an Clustern notwendig ist; stattdessen bleiben alle Teilanalysen im Ergebnis der Gesamt-Clusteranalyse bestehen, die somit

also mehr Informationen für eine explorative Datenanalyse bereitstellt (s. Manning & Schütze 1999: 495, 500). Außerdem lässt sich eine hierarchische Clusteranalyse graphisch als Dendrogramm darstellen (s. Abbildung 4.6), was diese Methode zusätzlich für die hier geplante Analyse textstruktureller Feature-Sets und Sequenzrepräsentationen geeignet macht.

4.2.2 Agglomeratives hierarchisches Clustering

Hierarchische Clusteringmethoden können weiter differenziert werden in **agglomerative** Methoden, die in einem *bottom-up*-Aufbau der Clustergruppen zunächst jedes Objekt als eigenen Cluster ansetzen, und **divisive** Methoden, die *top-down* vorgehen, also zu Beginn alle Datenpunkte als Teil eines einzigen Clusters ansetzen (s. Manning & Schütze 1999: 500f.). Da in der Fallstudie in Kapitel 6 agglomerative Clusteringmethoden zum Einsatz kommen, beschränkt sich die folgende Darstellung entsprechend auf diesen Typ.

	Text 1 ($\vec{d_1}$)	**Text 2** ($\vec{d_2}$)	**Text 3** ($\vec{d_3}$)
Text 2 ($\vec{d_2}$)	2.0		
Text 3 ($\vec{d_3}$)	3.16	3.74	
Text 4 ($\vec{d_4}$)	7.07	6.48	6.16

Tabelle 4.4: Distanzmatrix von Tabelle 4.1 (erweitert um Vektor $\vec{d_4}$)

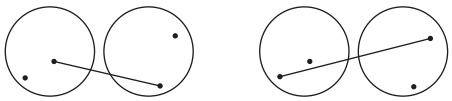


Abbildung 4.5: Agglomerationsmaße: Single-Linkage vs. Complete-Linkage (reproduziert nach Manning & Raghavan & Schütze 2009: 381, Abb. 17.3, Ausschnitt)

Grundlage der Berechnung von Clustern von Datenpunkten ist eine **Distanzmatrix**, die alle paarweisen Abstände der einzelnen Datenpunkte im Merkmalsraum gemäß einem zuvor gewählten Distanzmaß wie etwa dem euklidischen angibt. In der anschließenden Clusteranalyse werden dann diese Abstandsdaten zu den Datenpunkten verwendet, um Cluster als Mengen ähnlicher (d. h. im Merkmalsraum nahe beieinanderliegender) Datenpunkte zu bestimmen. Dazu werden beim agglomerativen Clustering die Distanzen zwischen vorhandenen Clustergruppen auf Grundlage der Abstände zwischen Gruppenmitgliedern berechnet und die naheliegendsten Cluster sukzessive vereinigt (gemergt). Es werden hier also zwei Distanztypen berechnet: zunächst die Distanzen im Vektorraum zwischen den Datenpunkten (Distanzmatrix) und dann, darauf basierend, im eigentlichen Clusteringverfahren die Distanzen zwischen Mengen dieser Datenpunkte.

Die Distanzen zwischen Mengen von Datenpunkten, gemäß denen diese zu Clustern verbunden werden, können dabei auf verschiedene Arten berechnet werden (vgl. Abbildung 4.5): Die einfachsten dieser sog. **Agglomerationsmaße** (auch: Linkage-Typen) sind vom Typ **Single-Linkage**, bei dem der Abstand zwischen zwei Clustern über den Abstand ihrer beiden naheliegendsten Elemente definiert ist („maxi-

mum similarity", Manning & Raghavan & Schütze 2009: 381f.), und **Complete-Linkage**, bei dem der Abstand zwischen zwei Clustern über den Abstand ihrer beiden am weitesten voneinander entfernten Elemente definiert ist („minimum similarity", Manning & Raghavan & Schütze 2009: 381f.). Daneben gibt es weitere Agglomerationsmaße wie Average-Linkage, das nicht nur die Abstände zwischen einzelnen Datenpunkten in den Clustern berücksichtigt, sondern zwischen allen Datenpunkten in den Clustern (Manning & Raghavan & Schütze 2009: 388f.), und so mögliche Probleme der beiden erstgenannten Linkage-Typen vermeiden kann (insbesondere Chaining bei Single-Linkage, s. u.; bei Complete-Linkage kann das Ergebnis ggf. durch Ausreißer gestört werden, s. Manning & Raghavan & Schütze 2009: 382). Außerdem ist hier Wards Agglomerationsmethode (Ward 1963; vgl. Manning & Raghavan & Schütze 2009: 399) zu nennen, bei der Cluster nach dem Kriterium minimaler Varianz zwischen den Gruppen fusioniert werden (s. Murtagh & Legendre 2014).[5]

Die Wahl der Agglomerationsmethode ist dabei maßgeblich für das Clusteringergebnis (s. James u. a. 2017: 396f.; vgl. auch 4.2.3 zur Cluster-Evaluation), denn je nach Methode (und auch in Abhängigkeit vom primären Distanzmaß für die Distanzmatrix) erhält man auf denselben Daten unterschiedliche Clustering-Resultate, die abhängig von dem Datentyp und der Fragestellung vorteilhaft oder nachteilig sein können:

> Ward's minimum variance method aims at finding compact, spherical clusters. The complete linkage method finds similar clusters. The single linkage method (which is closely related to the minimal spanning tree) adopts a 'friends of friends' clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods. (R Core Team 2020)

Als lokales Maß, das Cluster aufgrund ihrer nächstliegenden Elemente verbindet, produziert Single-Linkage langgezogene Cluster (sog. Chaining), was häufig unerwünscht ist (s. Manning & Raghavan & Schütze 2009: 382ff.). Bei Complete-Linkage werden dagegen die Clustergruppen basierend auf der Distanz ihrer entferntesten Mitglieder fusioniert (vgl. Abbildung 4.5): „This is equivalent to choosing the cluster pair whose merge has the smallest diameter" (Manning & Raghavan & Schütze 2009: 382). Als nicht-lokaler Abstandsvergleich („the entire structure of the clustering can influence merge decisions", Manning & Raghavan & Schütze 2009: 382) erzeugt Complete-Linkage entsprechend kompakte Cluster (Manning & Raghavan & Schütze 2009: 382) und verhindert Chaining, weswegen in der Fallstudie in Kapitel 6 Complete-Linkage als Standard-Agglomerationsmaß für Feature-Sets verwendet wird.

5 Für die Berechnung dieser Agglomerationsmaße sei auf die entsprechende Literatur verwiesen (Jain & Murty & Flynn 1999: 275ff.; Manning & Schütze 1999: 503ff.; Manning & Raghavan & Schütze 2009: 382ff.).

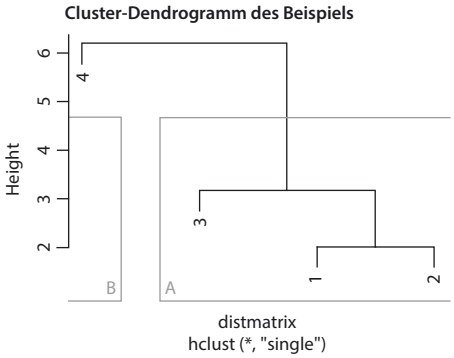


Abbildung 4.6: Beispiel Cluster-Dendrogramm

$(\vec{d_1}, \vec{d_2})$	$(\vec{d_3}, \vec{d_1})$	$(\vec{d_4}, \vec{d_3})$
2.0	3.16	6.16

Tabelle 4.5: Distanzen im Single-Linkage-Beispiel

Für folgendes Beispiel einer agglomerativen Clusteringanalyse wird allerdings der Single-Linkage-Clusteringalgorithmus als einfachste Agglomerationsmethode vorgestellt; für das Beispiel wird das Feature-Set aus Tabelle 4.1 um einen Vektor $\vec{d_4}$ = (5,5,5) erweitert (s. Tabelle 4.4). Als *bottom-up*-Methode setzt der agglomerative Clusteringalgorithmus (s. Manning & Raghavan & Schütze 2009: 378) zunächst jede Vektor-Textrepräsentation $\vec{d_i}$ (d. h. jeden Datenpunkt im Merkmalsraum) als eigenen Cluster an; schrittweise werden dann diese Mengen einzelner Datenpunkte gemäß der Distanz zwischen diesen Clustern zu größeren Clustergruppen vereinigt. Gestartet wird dieses agglomerative Clustering im Beispiel mit den Clustern $\{\vec{d_1}\}$, $\{\vec{d_2}\}$, $\{\vec{d_3}\}$, $\{\vec{d_4}\}$. Zunächst werden $\{\vec{d_1}\}$ und $\{\vec{d_2}\}$ als ähnlichste Cluster gemäß dem Single-Linkage-Maß als die Objekte mit dem geringsten Abstand gemergt; Resultat ist ein Cluster $\{\vec{d_1}, \vec{d_2}\}$. Es gibt damit drei Cluster: $\{\vec{d_1}, \vec{d_2}\}$, $\{\vec{d_3}\}$, $\{\vec{d_4}\}$. Da die Vektoren $\vec{d_3}$ und $\vec{d_1}$ den nächstkleineren Abstand haben, wird $\{\vec{d_3}\}$ mit dem $\vec{d_1}$-enthaltenden Cluster gemergt und es gibt nun zwei Cluster, nämlich Cluster A: $\{\vec{d_1}, \vec{d_2}, \vec{d_3}\}$ sowie Cluster B: $\{\vec{d_4}\}$ (s. Abbildung 4.6). An dieser Stelle beendet der Algorithmus das Gruppieren, da bereits die minimale Anzahl an Clustern erreicht ist (zwei Cluster).

Das Resultat einer hierarchischen Clusteranalyse kann über ein sog. **Dendrogramm** (s. Abbildung 4.6) visualisiert werden, d. h. über eine graphische Darstellung der Cluster-Abstände als Baumstruktur: Die Höhen, an denen die Cluster verzweigen, entsprechen dabei den durch den Agglomerationsalgorithmus berechneten Distanzen zwischen den Gruppen (vgl. Height-Achse in Abbildung 4.6). Beim Single-Linkage-Clustering lassen sich diese direkt aus der Distanzmatrix der Vektoren als die jeweils kürzeste Distanz zwischen den Vektoren zweier Cluster ablesen (s. Tabelle 4.5). Beispielsweise entspricht gemäß Single-Linkage-Agglomerationsmaß die Distanz zwischen dem durch den Vektor $\vec{d_4}$ definierten Datenpunkt und dem von den übrigen Datenpunkten gebildeten Cluster dem Abstand zu $\vec{d_3}$ (als dem Vektor mit dem kürzesten Abstand zu $\vec{d_4}$, vgl. Tabelle 4.4).

4.2.3 Evaluationsmethoden für Clusteringmodelle

Zur Evaluierung von Clusteringergebnissen gibt es eine Reihe statistischer Verfahren (s. Tan & Steinbach & Kumar 2006: 532ff.; Manning & Raghavan & Schütze 2009: 356ff.; Aggarwal 2015: 195ff.). Während Evaluationskriterien wie die Hopkins-Statistik oder Elbow- und Silhouette-Methoden die Güte der im Clustering gefundene Datenpartitionierung auf Grundlage der **internen** Strukturierung der Daten auswerten, messen **externe** Evaluationskriterien wie der Rand-Index den Grad der Übereinstimmung zwischen der Gruppierung des Clusteringresultats und vorliegenden Klassifizierungen der Daten (vgl. die Apriori-Texteinteilungen in 5.2.4), indem berechnet wird, wie viele Datenpunkte bzgl. einer solchen Vergleichspartitionierung vom Clusteringalgorithmus richtig gruppiert wurden. Eingesetzt werden Clusterevaluationsmethoden etwa zur Auswahl einer geeigneten Clusteringmethode in einem konkreten Anwendungsfall (insbesondere die externe Evaluation). Interne Cluster-Tendency-Maße geben in der Datenanalyse einen Hinweis darauf, inwieweit ein Feature-Set überhaupt Strukturierung aufweist oder welches die optimale Clusteranzahl ist (so z. B. in explorativen texttypologischen Untersuchungen wie bei Grzybek & Kelih & Stadlober 2005: 107).

4.2.3.1 Bestimmung der Cluster-Tendency

Zur Beurteilung, ob in einem Feature-Set überhaupt Ähnlichkeitsstrukturen zu entdecken sind, kann als Maß zur Beurteilung der sog. *Cluster-Tendency* die **Hopkins-Statistik** berechnet werden (s. Tan & Steinbach & Kumar 2006: 547f.; Aggarwal 2015: 157f.). Die Hopkins-Statistik bestimmt primär, ob die Daten gleichverteilt sind; dazu werden die Abstände der Datenpunkte eines Samples von p Datensätzen (w_i) zu den nächsten Nachbarn des Datensatzes mit den Abständen von p Datenpunkten eines entsprechend generierten Zufallssamples im Merkmalsraum (u_i) zu den nächsten Nachbarn des Datensatzes verglichen (s. Aggarwal 2015: 158):

$$\text{Hopkins statistic } H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i} \tag{4.4}$$

Je stärker der im Intervall $[0; 1]$ befindliche Wert sich von 0.5 (d. h. der Gleichverteilung) unterscheidet, desto höher ist die Cluster-Tendency der Daten.[6] Durch Bestimmung und Vergleich der Hopkins-Werte von Feature-Subsets kann im Rahmen einer Clustering-Mustererkennungsaufgabe auch eine geeignete Auswahl der zu berück-

6 In der hier verwendeten Version dieser Statistik deutet ein Wert deutlich unter 0.5 auf auffindbare Strukturen hin, da in diesem Fall die w_i-Abstandswerte des Datensatzsamples niedriger sind als die u_i-Abstandswerte des Zufallssamples; häufig wird stattdessen auch der Wert 1-H berechnet, dann sollte der Wert entsprechend deutlich oberhalb von 0.5 liegen, vgl. Aggarwal 2015: 158.

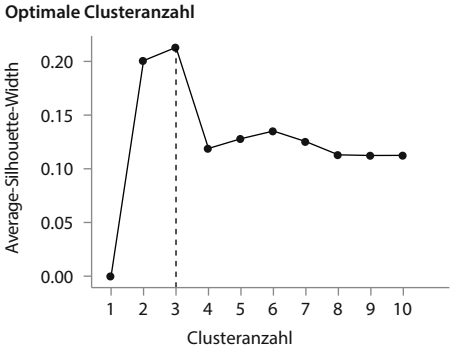
sichtigenden Merkmale eines Feature-Sets im Sinne einer Feature-Selection getroffen werden (s. Aggarwal 2015: 154ff.).

4.2.3.2 Methoden zur Bestimmung der optimalen Clusteranzahl

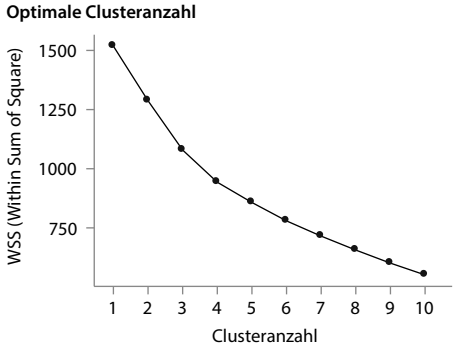
Eine Methode zur Bestimmung der optimalen Clusteranzahl ist die sog. **Elbow**-Methode, die das Mittel der Fehlerquadrate für die verschiedenen Clusteranzahlen berechnet (Aggarwal 2015: 196; vgl. Grzybek & Kelih & Stadlober 2005: 107 als Texttypologie-Methode). Die optimale Clusteranzahl kann dann bestimmt werden als die Stelle, an der der gegen die Clusteranzahl geplottete Wert abknickt; dieser ‚Elbow'-Wert ist die Clusteranzahl, ab der die Varianz innerhalb der Cluster – d. h. die Streuung um den Cluster-Mittelwert – durch weitere Partitionierung nicht mehr bedeutend abnimmt (Aggarwal 2015: 198; vgl. James u. a. 2017: 386f.).

Eine alternative Methode zur Bestimmung der optimalen Clusteranzahl ist die Bestimmung des Maximums der sog. **Average-Silhouette-Width** (ASW, Rousseeuw 1987; vgl. Aggarwal 2015: 196f.): „Each cluster is represented by a so-called *silhouette*, which is based on the comparison of its tightness and separation" (Rousseeuw 1987: 53, Hervorhebung im Original). Die ASW kann auch als weiteres Cluster-Qualitätsmaß (neben der Hopkins-Statistik) herangezogen werden, insbesondere bei Sequenzdaten. Der Maximalwert 1 impliziert eine sehr starke Cluster-Tendenz (s. Rousseeuw 1987: 60).

Die Festlegung der Anzahl der Clustergruppen ist ein zentraler Schritt in der Analyse des Resultats einer hierarchischen Clusteranalyse und insbesondere auch für die in dieser Arbeit angestrebte Erstellung einer Textstrukturtypologie relevant für die anschließenden Analysen (vgl. auch Grzybek & Kelih & Stadlober 2005: 107).



Plot 4.2.1: Silhouette-Plot (Beispiel)



Plot 4.2.2: Elbow-Plot (Beispiel)

4.2.3.3 Vergleich von Clusteringergebnissen

Für die Beurteilung von Clusteringergebnissen in Abhängigkeit von externen Kriterien (also einer a priori gegebenen Gruppeneinteilung) können verschiedene Maße herangezogen werden, u. a. Informationsmaße wie *Purity*, Gini-Index oder *Mutual Information* sowie der hier vorgestellte Rand-Index als ‚Entscheidungsmaß' (s. Manning & Raghavan & Schütze 2009: 356ff.; Aggarwal 2015: 198f.).

Der **Rand-Index** kann grundsätzlich für Vergleiche zweier Gruppierungen eingesetzt werden, also sowohl zum Vergleich zweier Clusteringergebnisse (unterschiedlicher Modelle auf denselben Daten) als auch als Evaluationsmaß eines Clusteringergebnisses auf Grundlage einer gegebenen Apriori-Gruppierung, d. h. als externes Kriterium.[7] Der Rand-Index misst den Anteil der korrekt vorgenommenen Gruppierungen zweier Objekte zu einem Cluster und damit die Accuracy des Clusteringresultats (Manning & Raghavan & Schütze 2009: 359; zu Accuracy s. 4.3.3). Der hier verwendete **Adjusted-Rand-Index** korrigiert den Rand-Index bei ungleichen Gruppengrößen (s. Manning & Raghavan & Schütze 2009: 373).

4.2.3.4 Feature-Importance für Clustering-Gruppen

Durch Verwendung der in einer Clusteranalyse gefundenen Clustergruppierungen als Response-Klassenlabels in einer Klassifikationsaufgabe (vgl. Ismaili & Lemaire & Cornuéjols 2014) kann für das im Clustering verwendete Feature-Set ein Ranking seiner Merkmale erstellt werden, indem sog. **Feature-Importance**-Werte für die Differenzierung der Clustergruppen berechnet werden (Details s. 4.3.3):[8]

> The objective of this work is to propose a simple way to identify the most relevant features from the output of a clustering. In order to retain all variables, we rank the variables according to their importance without doing a selection. The main idea is to turn this

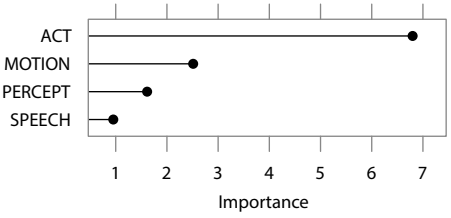
7 Zur Berechnung s. Manning & Raghavan & Schütze 2009: 359 (Hervorhebung im Original): „clustering [...] as a series of decisions, one for each of the $N(N-1)/2$ pairs of documents in the collection. We want to assign two documents to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A false positive (FP) decision assigns two dissimilar documents to the same cluster. A false negative (FN) decision assigns two similar documents to different clusters. The *Rand index* (RI) measures the percentage of decisions that are correct."

$$\text{Rand Index} = \frac{TP + TN}{TP + FP + TP + FN} \tag{4.5}$$

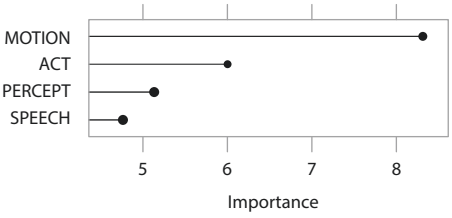
8 Durch den in dieser Arbeit verwendeten Random-Forest-Klassifikator ist eine direkte Berechnung der Importance-Werte der Variablen aus einem das gesamte Feature-Set enthaltenden Klassifikationsmodell möglich, ohne (wie bei Ismaili & Lemaire & Cornuéjols 2014: 161) für jede Variable einen eigenen Klassifikator trainieren und die Accuracy vergleichen zu müssen.

> problem into a supervised classification problem where the cluster membership (ID-cluster) is used as a target class. Then, for each variable, we use a supervised classification algorithm to predict the ID-cluster. (Ismaili & Lemaire & Cornuéjols 2014: 160f.)

In der Auswertung in Kapitel 6 können diese Clustering-bezogenen Feature-Evaluationsdaten direkt verglichen werden mit den entsprechenden Feature-Evaluationsdaten der Klassifikation mit Apriori-Klassen. So sieht man z. B. in Plots 4.2.3 und 4.2.4, dass die mit Random-Forest-Klassifikator berechneten Feature-Importance-Werte des Clusterings für die Ereignistypik-Parametrisierung den Werten der `GENRE`-Apriori-Textsorteneinteilung (s. 5.2.3) ähneln, nämlich mit `MOTION` und `ACTION` als wichtigsten differenzierenden Merkmalen (allerdings mit vertauschter Reihenfolge im Ranking) – und im Gegensatz zu den beiden anderen Apriori-Kategorisierungen `COMM_SIT` und `DISC_STRUCT`, bei denen sich in der Feature-Ranking-Analyse jeweils `SPEECH` als *most important feature* für die Klassendifferenzierung zeigt (vgl. 6.4.1).



Plot 4.2.3: Feature-Importance für Clustergruppen im Ereignistypik-Feature-Set (Beispiel)



Plot 4.2.4: Feature-Importance für GENRE-Klassen im Ereignistypik-Feature-Set (Beispiel)

4.2.4 Visualisierung und Analyse von Clusteringergebnissen

Neben der Analyse durch statistische Evaluationsmetriken kann eine gefundene Partitionierung im n-dimensionalen Feature-Raum auch durch graphische Darstellungstechniken visuell inspiziert werden, insbesondere über Methoden der Dimensionsreduktion, die die Struktur der n-dimensionalen Daten auf wenige darstellbare Dimensionen projizieren, oder auch über verschiedene Darstellungsmethoden der deskriptiven Statistik, z. B. über Plots von gruppenspezifisch differenzierten Durchschnittswerten.

4.2.4.1 Parallelkoordinatenplot und Streudiagramm

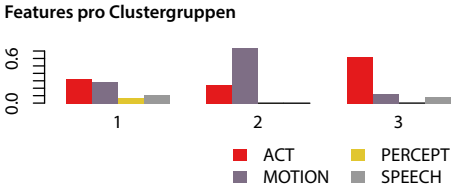
Ein Parallelkoordinatenplot stellt alle Datenpunkte eines geclusterten Feature-Sets als verbundene Linien in einem Koordinatensystem dar, dessen parallel angeordnete

Achsen jeweils einer Feature-Dimension entsprechen, wobei die Gruppenzugehörigkeit farblich markiert wird (s. Plot 3.8.1 für ein Beispiel).

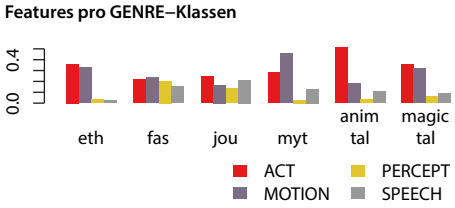
Einzelne Merkmale (insbesondere solche, die sich im Feature-Selection-Prozess als gruppendifferenzierend herausstellen) können mit paarweisen Streudiagrammen (Scatterplots) für eine weiterführende Feature-Analyse aufgetragen werden. Clustergruppen- bzw. Textsorten-Klassenzugehörigkeit kann dabei farblich differenziert dargestellt werden (s. Plot 4.3.1 für ein Beispiel).

4.2.4.2 Gruppierte Average-Scores-Barplots

Mit gruppierten Average-Scores-Barplots können für die Merkmale eines Feature-Sets die Durchschnittswerte in verschiedenen Datengruppen dargestellt werden. Mit dieser Visualisierungsmethode lassen sich sowohl die Werteverteilungen über Clustergruppen (vgl. Plot 4.2.5) als auch die über Klassen verschiedener Apriori-Textsortenkategorisierungen (vgl. Plot 4.2.6) darstellen. Dazu werden die Daten gemäß der jeweiligen Gruppeneinteilung aufgesplittet und die gruppenspezifischen Durchschnittswerte berechnet.



Plot 4.2.5: Clustergruppierte Average-Scores-Barplots des Ereignistypik-Feature-Sets (Beispiel)



Plot 4.2.6: GENRE-gruppierte Average-Scores-Barplots des Ereignistypik-Feature-Sets (Beispiel)

4.2.4.3 PCA-Clusterplots

Die Hauptkomponentenanalyse (PCA = Principal Component Analysis) ist eine **Dimensionsreduktionsmethode**, die zur Visualisierung von Clustertypologien eingesetzt werden kann (s. Beyerer & Richter & Nagel 2017: 57). Dabei wird der Vektorraum so transformiert, dass wenige Dimensionen (die sog. Hauptkomponenten) als Achsen des rotierten Vektorraums die Hauptinformation tragen (**Feature-Projection**). Dies sind die Dimensionen größter Varianz (Streuung) im Vektorraum (s. James u. a. 2017: 231, 376); die Hauptkomponenten sind dabei die linearen Kombinationen der Features (X_1, X_2, X_3...), die die Varianz der Daten maximieren; die erste Hauptkomponente ist z. B. $\phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{n1}X_n$; die folgenden Hauptkomponenten (zweite usw.) sind auf die vorherigen jeweils orthogonal (s. James u. a. 2017: 231, 375f.).

Die Achse der ersten Hauptkomponente gibt also die Richtung der größten Streuung der Daten an; diese Achse ist gleichzeitig die Gerade, die so nah wie möglich an den Daten liegt, d. h. die Gerade durch den Datenmittelpunkt, sodass die Summe des Abstands der Daten dazu minimal wird (James u. a. 2017: 231). Die ersten beiden Hauptkomponenten können für eine zweidimensionale Projektion des ursprünglichen, höherdimensionalen Vektorraums verwendet werden, um die Güte der Trennung der Gruppen visuell zu analysieren (vgl. Manning & Schütze 1999: 527; James u. a. 2017: 375ff.).

4.3 Klassifikationsmethoden

4.3.1 Klassifikation als überwachte Klassifizierung

Als explorative Klassifizierungsmethode, in der die Datenpunkte eines Feature-Sets aufgrund ihrer Lage im Merkmalsraum a posteriori in Klassen eingeteilt werden, geht das eben besprochene Clustering grundsätzlich *strukturentdeckend* vor: Durch das Auffinden von Mustern aufgrund von Ähnlichkeitsbeziehungen in den Daten ermöglichen Clusteringverfahren somit eine datengestützte, empirisch-induktive Generierung von Texttypologien. Anders der zweite Typ von Klassifizierungsmethoden, die in dieser Arbeit für eine *strukturprüfende* Analyse von textstrukturellen Feature-Sets in Abhängigkeit von a priori gegebenen Genre-Klassifizierungen herangezogen werden: Die sog. **Klassifikationsmethoden** des Machine-Learnings setzen dabei – im Gegensatz zum Clustering – eine Einteilung der zu untersuchenden Objekte in Klassen voraus und lernen darauf aufbauend die Abbildung der Objektrepräsentationen auf diese vorgegebene Klasseneinteilung anhand ihrer Feature-Werte. In der Klassifikation als überwachter Klassifizierung (***supervised***) werden Objekte also nicht datenbasiert (auf Grundlage ihrer Merkmalsausprägungen) eingeteilt, sondern es wird eine Funktion gelernt – ein sog. **Klassifikator** (s. Manning & Schütze 1999: 575f.; Manning & Raghavan & Schütze 2009: 256), der angibt, wie Objekte aufgrund ihrer Merkmalsausprägungen in bereits vorgegebene Gruppen eingeteilt werden können. Mit dieser Methodik lässt sich u. a. überprüfen, ob (bzw. zu welchem Grad) sich theoretisch angenommene Typologien tatsächlich in den Daten abbilden; die Klassifikation ermöglicht damit im Rahmen dieser Arbeit eine Analyse textstruktureller Parameter in Abhängigkeit verschiedener Apriori-Kategorisierungen von Texten (vgl. 5.2.4).[9]

9 Gemäß der dieser Arbeit zugrunde liegenden kognitiven These der Textverarbeitung über Text-Weltmodelle (s. Kapitel 1) kann man annehmen, dass auch die Kognition eine solche Klassifikation im Sinne einer überwachten Klassifizierung aufgrund von in der Erfahrung gegebenen Klassen vornimmt, indem neue Situationen aufgrund von Strukturähnlichkeit bestimmten, zuvor durch Typisierung vergangener Situationen gewonnenen Situationsklassen zugeordnet werden, wodurch das damit verbundene schematische Weltmodell zur Verarbeitung der Situation aktiviert wird; gleichzeitig dient dann jede neue Situation als Trainingsinstanz zur Verfeinerung des auf diese Weise induktiv gelernten TWM-Klassifikationsmodells.

4.3.1.1 Vorgehen Klassifikation

Zum Aufbau eines Klassifikationsmodells wird jede Textrepräsentation $\vec{d_i}$ der sog. Trainingsmenge mit einem Klassenlabel c (Response) versehen und zu einem Tupel $(\vec{d_i}, c_i)$ vereint; die Feature-Set-Datenmatrix wird also um eine Spalte für das Klassenlabel als abhängige Variable erweitert:

	Merkmal A (f_1)	**Merkmal B** (f_2)	**Merkmal C** (f_3)	**Klassenlabel** c
Text 1 ($\vec{d_1}$)	1	2	0	Klasse-A
Text 2 ($\vec{d_2}$)	1	4	0	Klasse-B
Text 3 ($\vec{d_3}$)	0	2	3	Klasse-A

Tabelle 4.6: Um Klassenlabel erweiterte Datenmatrix für Klassifikation

Anschließend wird ein **Klassifikationsalgorithmus** auf die Menge dieser Trainingsinstanzen angewendet; dabei wird eine Funktion gelernt, die jedes $\vec{d_i}$ auf seine Klasse c_i abbildet. Diese Funktion kann dann als Klassifikator verwendet werden, um neue Objekte $\vec{d_x}$ auf eine Klasse abzubilden. Die Güte eines solchen Klassifikationsmodells kann anhand einer Testmenge von gelabelten Instanzen der Form $(\vec{d_i}, c_i)$ überprüft werden (s. 4.3.3 zur Evaluation von Klassifikatoren).

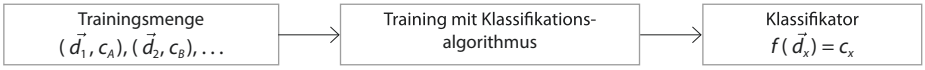


Abbildung 4.7: Ablaufdiagramm des Trainings eines Klassifikators

Im Gegensatz zu der unüberwachten Klassifizierung des Clusterings, bei der die Klassen automatisch aus den Daten gelernt werden, sind beim überwachten Lernen die Klassen vorgegeben: „a supervisor (the human who defines the classes and labels training documents) serves as a teacher directing the learning process“ (Manning & Raghavan & Schütze 2009: 256). Gelernt wird im Trainingsprozess also die Abbildung der Objekte auf diese vorgegebenen Klassenlabels. Das trainierte Klassifikationsmodell kann dann zur Vorhersage des Klassenlabels (*prediction*) noch nicht kategorisierter Objekte verwendet werden, indem der Klassifikator (die Klassifikationsfunktion) die Objektrepräsentationen als Input nimmt und diese – der aus den Trainingsdaten gelernten Zuordnung entsprechend – auf die Klassenlabels abbildet.

In dieser Arbeit werden als Klasseneinteilungen für eine Verwendung in klassifikationsbasierten Feature-Analysen vier verschiedene Typen von Apriori-Genre-Kategorisierungen herangezogen (u. a. zwei Textsorteneinteilungen, s. 5.2.4), um die auf den Korpusdaten trainierten Klassifikationsmodelle für diese verschiedenen Genre-Einteilungen (insbesondere deren Feature-Importance-Daten) miteinander sowie mit

den Ergebnissen der datengestützten, automatischen Klassifizierung durch die induktive Clusteranalyse im Sinne einer explorativen Feature-Analyse zu vergleichen.[10]

4.3.1.2 Klassifikationsalgorithmen

Grundsätzlich richtet sich die Auswahl eines Algorithmus für eine Klassifikationsaufgabe zum einen nach dem Skalenniveau der Daten (metrisch, kategorial, gemischt) – sowohl nach dem der Features, der sog. Prädikatorvariablen, als auch nach dem der vorherzusagenden, abhängigen Variablen (Klassen-Response) –, zum anderen nach Kriterien der Performanz sowie nach dem Anwendungsbezug. Da hier Textklassen – d. h. also kategoriale Labels – vorhergesagt werden sollen, kommen **Klassifikationsmethoden** im engeren Sinn zum Einsatz, im Gegensatz zu Regressionsmethoden für die Vorhersage von kontinuierlichen, metrischen Variablen. Da als Repräsentationsformat für die Text-Objekte hier das Datenrepräsentationsmodell verwendet wird, sind insbesondere Methoden der Vektorraum-Klassifikation relevant (s. Manning & Raghavan & Schütze 2009: 289ff.; Manning & Schütze 1999: 577f.).

Es existiert eine Vielzahl unterschiedlicher Typen von Klassifikationsalgorithmen, die in diesem Bereich der Textklassifikation Anwendung finden (s. Manning & Raghavan & Schütze 2009: 289ff.; Manning & Schütze 1999: 607f.) und auch hier im Bereich der textstrukturellen Klassifizierung zum Einsatz kommen können.[11] In konkreten Anwendungen von Genre-Klassifikation (s. Abschnitt 2.3) kommt beispielsweise bei Lapshinova-Koltunski & Zampieri 2018 ein auf N-Gramme angewendeter Naive-Bayes-Klassifikator zum Einsatz, bei Grzybek & Kelih & Stadlober 2005 dagegen Diskriminanzanalysen, angewendet auf Feature-Sets von Wortlängen-Maßen.

In dem hier geplanten Pretest (Kapitel 6) sollen Klassifikationsmodelle als Analysemodelle auf einem kleinen Testkorpus zur Untersuchung der in Kapitel 3 erarbeiteten Parameter einer kognitiven Texttypologie eingesetzt werden. Statt der Entwicklung eines Modells für die Vorhersage der Genrezugehörigkeit unbekannter Texte sollen hier also die Klassifikationsmodell-Parameter evaluiert werden, d. h. vor allem eine Gewichtung der Parameter durch **Feature-Selection**-Methoden erreicht werden.

Mit **Random-Forest** wird in dieser Arbeit für die Klassifikation der Feature-Set-basierten Textstruktur-Parameter eine Methode gewählt, die sich einerseits als Analysemodell zur Feature-Evaluation eignet, da sich aus diesen Entscheidungsbaum-

10 Vgl. das Vorgehen bei Grzybek & Kelih & Stadlober 2005, die im Rahmen einer auf klassischen Maßen statistischer Linguistik basierenden texttypologischen Untersuchung mit Linearer Diskriminanzanalyse als Klassifikationsmethode die Vorhersagekraft verschiedener Apriori-Text-Klassifizierungen vergleichen (z. B. Funktionalstil vs. Diskurstypen, s. Grzybek & Kelih & Stadlober 2005: 118).

11 Für eine allgemeine Darstellung der verschiedenen Typen von Klassifikationsalgorithmen sei hier auf die Einführungen bei Manning & Schütze 1999, Manning & Raghavan & Schütze 2009, James u. a. 2017, Hastie & Tibshirani & Friedman 2009 sowie Beyerer & Richter & Nagel 2017 verwiesen.

basierten Modellen direkte Feature-Importance-Metriken ergeben (Géron 2017: 192; James u. a. 2017: 319). Andererseits ist das Random-Forest-Klassifikationsverfahren als Erweiterung des einfachen Entscheidungsbaum-Ansatzes durch das Mitteln über eine Vielzahl von Entscheidungsbäumen sowie die Randomisierung und das Sampling sowohl der Objekte als auch der Features der Trainingsdaten eine in der Anwendung robuste, vielseitig einsetzbare Methode, die sich dadurch auch für kleinere Datensätze wie für den des Pretests dieser vorliegenden Arbeit eignet (vgl. Géron 2017: 187; Beyerer & Richter & Nagel 2017: 226).[12]

Im Folgenden werden die Random-Forest-Klassifikationsmethode vorgestellt und die Möglichkeiten für eine darauf basierende Feature-Analyse und -Evaluation aufgezeigt. (Klassifikationsmethoden für sequentielle Daten werden im anschließenden Abschnitt 4.4 besprochen.)

4.3.2 Klassifikation mit Ensemblemethoden (Random-Forest)

Die **Random-Forest**-Klassifikationsmethode (entwickelt von Breiman 2001) ist eine Erweiterung des Entscheidungsbaum-Verfahrens, basierend auf dem sog. **Bagging**-Ansatz (Breiman 1996), bei dem mehrere alternative Entscheidungsbäume für dasselbe Klassifikationsproblem erzeugt werden und (im Regressionsfall) der Durchschnitt der Vorhersagen berechnet bzw. (im Klassifikationsfall) eine Mehrheitswahl aus den Vorhersagen von diesem sog. **Ensemble** von Bäumen bestimmt wird (James u. a. 2017: 316f.; Beyerer & Richter & Nagel 2017: 222). Random-Forest erweitert dabei diese Bagging-Methode, indem bei der Erzeugung des Entscheidungsbaum-Ensembles ein doppelter Einsatz von **Randomisierung** zum Tragen kommt (vgl. Beyerer & Richter & Nagel 2017: 223) – daher auch die Bezeichnung Random-Forest (randomisierter ‚Wald', also Randomisierung eines Ensembles von Entscheidungsbäumen). Solche randomisierten Ensemble-Klassifikatoren verhindern ein Overfitting (s. Beyerer & Richter & Nagel 2017: 222), d. h. eine Überanpassung des Resultats des Lernalgorithmus an die Daten, wenn viele Features bzw. wenige Datensätze verwendet werden, was zu einem schlechteren Vorhersageergebnis auf unbekannten Daten führt; indem solche Ensemble-Klassifikatoren also über die Kombination vieler Einzelklassifikatoren Varianz reduzieren (James u. a. 2017: 316), erzielen sie bessere Resultate als einzelne Entscheidungsbäume.

4.3.2.1 Entscheidungsbäume

Die Klassifikation mit **Entscheidungsbäumen** (Manning & Schütze 1999: 578ff.; s. auch Beyerer & Richter & Nagel 2017: 215ff.) basiert auf schrittweisen Partitionie-

12 Random-Forest kann außerdem sowohl für Regression als auch Klassifikation eingesetzt werden, vgl. Beyerer & Richter & Nagel 2017: 226: „Since they build on decision trees, random forests can be used with features on all measurement scales and mixtures thereof."

rungen des Vektorraums (James u. a. 2017: 303ff.) in je zwei Regionen (Halbebenen, vgl. James u. a. 2017: 307). Jede dieser Partitionen ist definiert durch eine auf die Achse der Feature-Dimension eines Features X_i orthogonale Gerade, die durch ein sog. Split-Kriterium t definiert wird ($X_i \leq t$); diese Gerade teilt die Datenpunkte also nach ihrer Werteverteilung bzgl. Merkmal X_i in zwei Mengen ein, je nachdem, ob für den X_i-Feature-Wert für ein Objekt $X_i \leq t$ gilt oder nicht.

Man betrachtet nun die vorherzusagenden Klassen-Werte (*prediction*) der Trainingsobjekte in den beiden durch den Split entstehenden Regionen und weist jeder Region die dort am häufigsten vorkommende Klasse zu: „assign an observation in a given region to the *most commonly occurring class* of training observations in that region" (James u. a. 2017: 311). Das **Splitting** wird rekursiv wiederholt, bis ein Stop-Kriterium erreicht ist, also z. B. nur noch x Objekte pro Region enthalten sind. Am Ende ist der Vektorraum in verschiedene Regionen eingeteilt (vgl. Abbildung 4.8), deren Trennung durch die Splits festgelegt ist (s. James u. a. 2017: 306). Durch die rekursive Anwendung von Split-Regeln – das sog. rekursive binäre Splitting des Vektorraums (s. James u. a. 2017: 306; Beyerer & Richter & Nagel 2017: 218; Manning & Schütze 1999: 582f.) – entsteht eine hierarchische Baumstruktur:

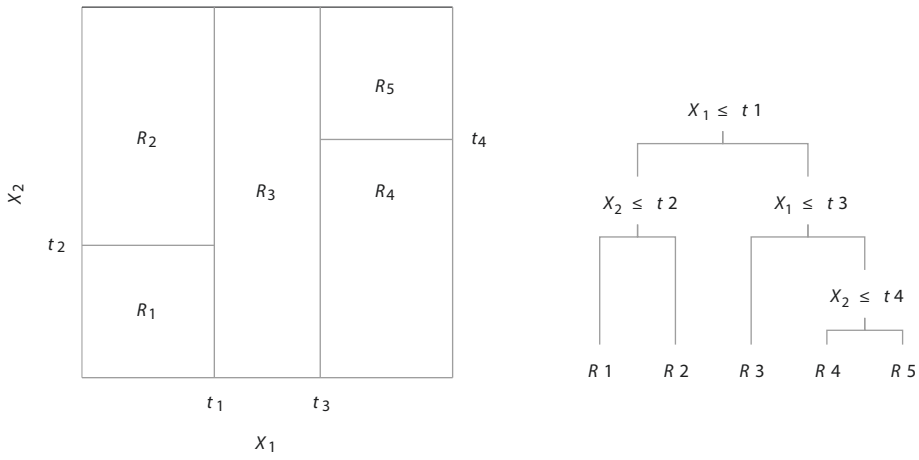


Abbildung 4.8: Rekursives binäres Splitting des Vektorraums, Beispiel mit zwei Prädikator-Merkmalen X_1 und X_2 (reproduziert nach James u. a. 2017: 308, Abb. 8.3)

An jedem Split-Knoten findet dabei eine binäre Entscheidung statt ($X_i \leq t$), welche die jeweilige Objektmenge in zwei Teilmengen aufteilt; entsprechend heißt die Struktur auch Entscheidungsbaum (*decision tree*); ein Blatt repräsentiert dabei den Wert der Klassen-Vorhersage für die entsprechende Region (d. h. die darin am häufigsten vertretene Klasse):

> The subset D_{Yes} is associated with the left branch of the split and D_{No} is associated with the right branch. On each branch, a node is again constructed according to the splitting criterion, but only using the samples that reach that node. This procedure is repeated recursively until a stopping criterion is met. (Beyerer & Richter & Nagel 2017: 218)

Das Training eines Entscheidungsbaumes geschieht also durch sukzessive Anwendung von **Split-** und **Stop-Kriterien** (Manning & Schütze 1999: 582f.; Beyerer & Richter & Nagel 2017: 218). Die Aufgabe des Klassifikationsalgorithmus besteht beim Training darin, bei jeder Partitionierung den Splitting-Wert t für das Feature X_i (Prädikatorvariable) zu finden (Split-Kriterium der Form $X_i \leq t$), bei welchem die Summe der Klassifikations-Fehlerraten („the fraction of the training observations in that region that do not belong to the most common class", James u. a. 2017: 311) für beide durch den Split erzeugten Regionen an dieser Stelle minimal wird (s. James u. a. 2017: 306f., 311f.):

$$\hat{p}_{mk} = \text{Anteil der Objekte der } k\text{-ten Klasse in der } m\text{-ten Region} \tag{4.6}$$

$$\text{Klassifikations-Fehlerrate} = 1 - \max_k(\hat{p}_{mk}) \tag{4.7}$$

Anstelle der Klassifikations-Fehlerrate werden als Kriterien für Entscheidungsbaum-Klassifikation häufig Informations-Maße wie der **Gini-Index** oder die Entropie verwendet (s. James u. a. 2017: 312):

$$\text{Gini-Index} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.8}$$

Als „measure of total variance across the K classes" (James u. a. 2017: 312) ist der Wert des Gini-Index klein, wenn alle $\hat{p}_{mk}$ annähernd 1 oder 0 sind (d. h. eine Klasse dominiert die Region, die Klassifikations-Fehlerrate ist also gering). Man spricht hier auch davon, dass der Split-Knoten mit niedrigem Gini-Index ‚rein' ist; dieser ist entsprechend ein Maß der sog. *node (im)purity*, vgl. James u. a. 2017: 312: „[…] a small value indicates that a node contains predominantly observations from a single class."

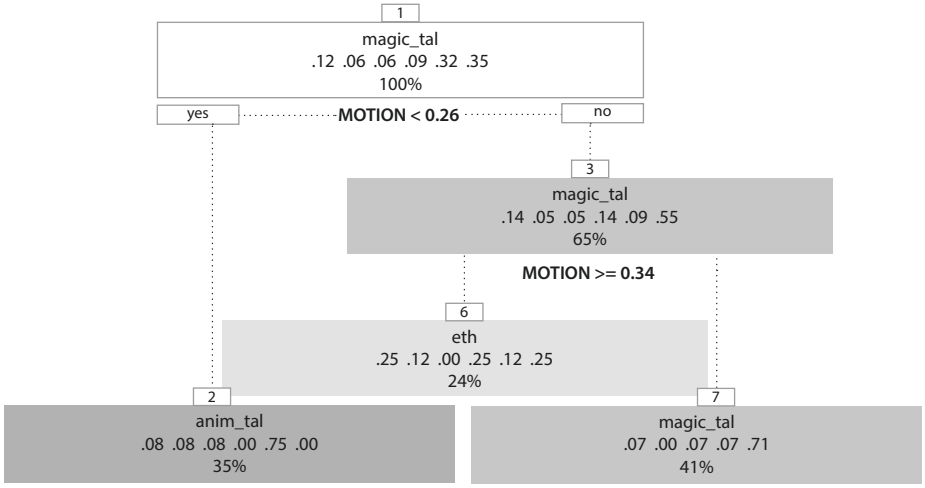
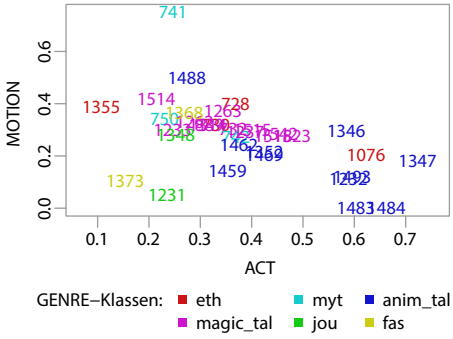


Abbildung 4.9: Entscheidungsbaum für die GENRE-Vorhersage (Ereignistypik)



Plot 4.3.1: Scatterplot zweier Ereignistypik-Features nach GENRE-Klassen (Beispiel)

Abbildung 4.9 gibt ein Beispiel für die Anwendung von Split-Kriterien: Die beiden Splits dieses Entscheidungsbaums zum Ereignistypik-Feature-Set des obugrischen Korpus (vgl. Plot 4.3.1) implizieren eine Aufteilung des Feature-Space bzgl. der MOTION-Dimension in drei Partitionen.

Entscheidungsbäume eignen sich gut als Analysemodelle, da dadurch ein für Menschen direkt lesbares Modell aus den Daten induziert wird (s. Géron 2017: 192; James u. a. 2017: 319):[13]

> The greatest advantage of decision trees is that they can be interpreted so easily. [...] This is not only invaluable in debugging one's own code and understanding a new problem domain, but it also allows one to explain the classifier to researchers and laymen alike, an important property in research collaboration and practical applications. (Manning & Schütze 1999: 588)

13 Entscheidungsbäume können auch als diskrete Regeln dargestellt werden, nämlich als Disjunktion der Pfade im Entscheidungsbaum (wobei ein Pfad die Konjunktion der einzelnen Entscheidungen ist, die in den Knoten getroffen werden), die zum selben Resultat, also übereinstimmender Vorhersage, führen (vgl. Mooney 2004: 381f.).

Auch werden Entscheidungsbäume als adäquatere Modelle der menschlichen Informationsverarbeitung und Handlungssteuerung gesehen als andere statistische Klassifikationsmodelle, s. James u. a. 2017: 315: „Some people believe that decision trees more closely mirror human decision-making than do the regression and classification approaches […]."

4.3.2.2 Bagging

Wie oben erwähnt, basiert die Random-Forest-Methode auf der Grundidee des **Bagging** (Breiman 1996; von *bootstrap aggregation*); dabei werden mehrere Entscheidungsbaum-Klassifikatoren kombiniert, indem das Trainingsset **gesampelt** wird (= Bootstrapping, s. James u. a. 2017: 187ff., 316f.). Ziel ist eine Reduktion der Varianz des zugrunde liegenden Klassifikationsmodells (James u. a. 2017: 316). Dadurch kann ein mögliches Overfitting verhindert und eine bessere Vorhersage erreicht werden: „However, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved" (James u. a. 2017: 316).

Die Klassifikation eines neuen Objekts geschieht dann durch Mehrheitswahl aus den verschiedenen Klassenvorhersagen, die aus der Eingabe der Feature-Repräsentation des Objekts in die Entscheidungsbäume des Ensembles resultieren (s. Hastie & Tibshirani & Friedman 2009: 588).

4.3.2.3 Random-Forest

Die Random-Forest-Klassifikation ist eine Erweiterung des Bagging-Ansatzes, in dem nicht nur die Trainingsdaten gesampelt werden (Bootstrapping), sondern zusätzlich jeweils ein Subsampling der Prädikator-Features vorgenommen wird:

> Random forests use randomization during training in two ways: First, each tree in the ensemble is trained on a random subsample of the training set. Second, at each node in each tree, only a random subspace of the feature space is considered for a split. (Beyerer & Richter & Nagel 2017: 223)

Bei jeder Berechnung eines Baumes für das Baum-Ensemble wird also jeweils nur eine Zufallsauswahl der Feature-Variablen berücksichtigt (für den Algorithmus s. Hastie & Tibshirani & Friedman 2009: 588; s. auch Tan & Steinbach & Kumar 2006: 276; Beyerer & Richter & Nagel 2017: 215ff.). Der Sinn des Feature-Space-Samplings besteht in der **Dekorrelation** der Bäume durch Einschränkung der beim Splitten berücksichtigten Prädikatoren (s. James u. a. 2017: 320; Hastie & Tibshirani & Friedman 2009: 597ff.), vgl. auch Beyerer & Richter & Nagel 2017: 226: „Bagging and random feature sub-sampling lead to better generalization properties compared to a single decision tree. This effect is commonly observed in ensemble methods."

Bei der Random-Forest-Klassifikation wird also pro Split nur eine begrenzte Anzahl m der p Features im Feature-Set der Trainingsdaten verwendet; üblicherweise werden $\sqrt{p}$ Features verwendet (s. Hastie & Tibshirani & Friedman 2009: 589).

Anzahl an Features pro Split: $m < p$ (typischerweise: $m = \sqrt{p}$) (4.9)

Die Anzahl an Features pro Split kann als sog. Tuning-Parameter verstanden werden, den es vor dem Training festzulegen gilt; hier können Methoden zum automatischen Auffinden des optimalen Wertes Anwendung finden (vgl. 6.1.4); vgl. Hastie & Tibshirani & Friedman 2009: 592: „In practice the best values for these parameters will depend on the problem, and they should be treated as tuning parameters."

Die Klassifikation eines neuen Objekts funktioniert wie bei den einfachen Bagging-Methoden über Mehrheitswahl aus den Klassifikationsergebnissen der Bäume des Random-Forest-Ensembles (s. Hastie & Tibshirani & Friedman 2009: 588).

4.3.3 Feature-Selection und Evaluation von Klassifikatoren

4.3.3.1 Feature-Selection und Feature-Evaluation

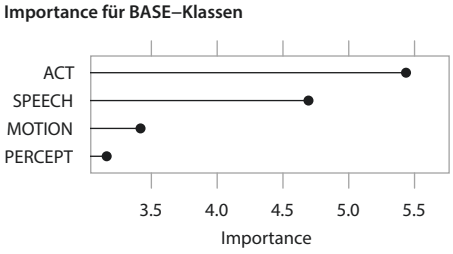
Unter **Feature-Selection** (Motoda & Liu 2002; Tan & Steinbach & Kumar 2006; Manning & Raghavan & Schütze 2009: 271ff.) versteht man allgemein das Auffinden der optimalen Menge an Merkmalen für eine Klassifikationsaufgabe, d. h. also die Kombination an Merkmalen, die als Trainingsinputdaten für ein Klassifikationsmodell das beste Resultat liefert; dazu gehört insbesondere die Eliminierung irrelevanter Features aus dem Feature-Space, vgl. Motoda & Liu 2002: 67: „Feature selection is a process that chooses a subset of M features from the original set of N features $(M \leq N)$, so that the feature space is optimally reduced according to a certain criterion."

Feature-Selection-Methoden evaluieren also die Features, die als Vorhersageparameter eines Klassifikationsmodells verwendet werden, hinsichtlich ihrer Vorhersagekraft, d. h. ihrer Wichtigkeit für die Diskrimination der Klassen. In dieser Arbeit, die eine Gewichtung der TWM-Parameter – also eine Untersuchung der Relevanz der in Kapitel 3 als Hypothesen vorgeschlagenen Parameter kognitiver Texttypen – anstrebt, werden entsprechend solche Feature-Selection-Verfahren als Methoden für diese angestrebte **Feature-Evaluation** verwendet. Mit der Entscheidungsbaum-basierten Random-Forest-Klassifikationsmethode kann eine sog. Embedded-Feature-Selection durchgeführt werden (Aggarwal 2015: 292), d. h. Feature-Set-Analyse-Kriterien können (s. Breiman 2001: 10; 23ff.) direkt aus dem Modell gewonnen werden:

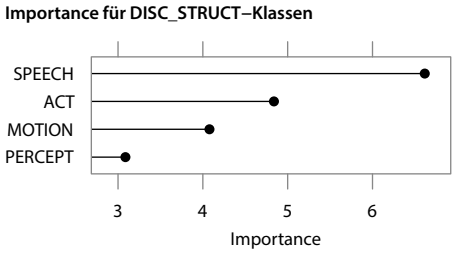
> Random forests can be used to reduce the dimensionality of the feature space by assessing each feature's importance using the left-out samples from bootstrapping and removing features with little importance (see Breiman[2001]: feature selection is embedded in a random forest [...]). (Beyerer & Richter & Nagel 2017: 226; vgl. Breiman 2001)

4.3.3.2 Feature-Importance

Die Relevanz der verschiedenen Merkmale eines Feature-Sets für die Klassendifferenzierung kann mit sog. **Feature-Importance**-Metriken bestimmt werden. Plot 4.3.2 und Plot 4.3.3 zeigen ein solches Ranking der Merkmale des Ereignistypik-Feature-Sets für zwei verschiedene Apriori-Textklassifizierungen (vgl. 5.2.4).



Plot 4.3.2: Feature-Importance für BASE-Klassen im Ereignistypik-Feature-Set (Beispiel)



Plot 4.3.3: Feature-Importance für DISC_STRUCT-Klassen im Ereignistypik-Feature-Set (Beispiel)

Als Maß für die Feature-Importance kann aus einem Random-Forest-Modell für jedes Feature die durchschnittliche **Abnahme des Gini-Index** über alle Entscheidungsbäume pro Split des jeweiligen Features gemessen werden: „[...] in the context of bagging classification trees, we can add up the total amount that the Gini index [...] is decreased by splits over a given predictor, averaged over all B trees" (James u. a. 2017: 319). Als Split-Kriterium ist dieser Index für die einzelnen Entscheidungsbäume direkt im Random-Forest-Modell gegeben (s. 4.3.2.1 zur Berechnung des optimalen Splitwertes über einen minimalen Gini-Index im Sinne einer minimalen Fehlerrate; vgl. auch Hastie & Tibshirani & Friedman 2009: 593ff.).

Als Maß der *node impurity* ist ein niedriger Wert des Gini-Index maßgeblich für die Entscheidung im Training des Random-Forest-Klassifikators, welches Feature jeweils gesplittet wird, denn ein minimaler Gini-Index-Wert minimiert entsprechend auch die Klassifikationsfehlerrate bei einer Entscheidung (Knoten im Entscheidungsbaum, s. 4.3.2.1). Bei jedem Split nimmt diese *node impurity* ab (s. Hastie & Tibshirani & Friedman 2009: 309; James u. a. 2017: 330; Breiman & Cutler 2020); die Klassifikation wird somit durch jede neue Partitionierung genauer. Je höher die durchschnittliche Abnahme für die Splits eines Features im Random-Forest-Ensemble also ist, desto wichtiger ist dieses Merkmal für die Differenzierung der Klassen, s. Breiman & Cutler 2020: „Every time a split of a node is made on variable m the gini

impurity criterion for the two descendent nodes is less than the parent node. Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance [...]."

4.3.3.3 Evaluationsmetriken für Klassifikationsmodelle

Für die Untersuchung der Vorhersagekraft von Klassifikationsmodellen gibt es eine Reihe an Standard-Evaluationsmaßen (s. Manning & Schütze 1999: 577; James u. a. 2017: 149), die die Genauigkeit der Vorhersage der auf einer Trainingsdatenmenge trainierten Klassifikatoren bzgl. einer Testmenge bestimmen (also auf neuen Daten beruhen, die nicht Teil des Trainings waren). Die Berechnung der **Evaluationsmetriken** wird im Folgenden beispielhaft für den einfachen Fall eines binären Klassifikators erläutert. Ein solcher Klassifikator sagt die Mitgliedschaft von Objekten in einer bestimmten Klasse voraus; es gibt hier zwei mögliche Vorhersagen und entsprechend zwei Klassenlabels: `JA` (Objekt ist enthalten) und `NEIN` (nicht enthalten). Die für die Berechnung der Maße notwendige Testmenge mit bekannter Klassenzuweisung der Objekte wird üblicherweise durch zufällige Partitionierung der Gesamtdaten vor dem Training in Trainings- und Testmenge erzeugt.[14]

	JA ist korrekt	**NEIN ist korrekt**
Vorhersage: JA	a (*true positive*-Vorhersage)	b (*false positive*-Vorhersage)
Vorhersage: NEIN	c (*false negative*-Vorhersage)	d (*true negative*-Vorhersage)

Tabelle 4.7: Kontingenztabelle für Evaluation eines binären Klassifikators (nach Manning & Schütze 1999: 577)

	Formel
Accuracy	$A = \frac{a+d}{a+b+c+d}$
Precision	$P = \frac{a}{a+b}$
Recall	$R = \frac{a}{a+c}$
F-score	$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}}$
Kappa	$\kappa = \frac{A - A_{expected}}{1 - A_{expected}} = 1 - \frac{1-A}{1-A_{expected}}$

Tabelle 4.8: Evaluationsmetriken für Klassifikatoren (s. Manning & Schütze 1999: 577; Carletta 1996: 252)

Tabelle 4.7 enthält alle vier möglichen Fälle der Übereinstimmung der Vorhersage durch den Klassifikator bei Anwendung auf der Testmenge und der in der Testmen-

14 Die Evaluationsmetriken können alternativ auch auf Grundlage von Resampling-Methoden wie Cross-Validation berechnet werden (s. 6.1.4).

ge tatsächlich gegebenen Klassenwerte; die Frequenzdaten (a, b, c, d) dieser Kreuztabelle sind dann Grundlage für die Berechnung der Evaluationsmaße in Tabelle 4.8. Folgende Metriken werden zur Evaluation von Klassifikationsmodellen verwendet:

- **Accuracy** misst den Anteil aller korrekten Vorhersagen (positive sowie negative);
- **Precision** misst den Anteil korrekter Vorhersagen an den positiven Vorhersagen (d. h. dass ein Objekt zur Klasse gehört), also wie viele der als Klassenmitglied vorhergesagten Elemente dies auch tatsächlich sind;
- **Recall** misst dagegen den Anteil der korrekten Vorhersagen bzgl. der Klassenmitglieder, also wie viele der Objekte, die als Mitglieder der Klasse vorhergesagt werden sollen, auch tatsächlich als solche vorhergesagt werden;
- der **F-score** ist eine gewichtete Kombination aus Precision und Recall;
- **Kappa**[15] ist die über die *expected accuracy* normalisierte, relative Accuracy und eignet sich zum Vergleich der Übereinstimmung von Klassifikationsmodellen bzgl. unterschiedlicher Klasseneinteilungen (Carletta 1996), da sowohl unbalancierte Klassenverteilungen (also z. B. 75% : 25% statt 50% : 50%) als auch eine unterschiedliche Klassenanzahl in den zu vergleichenden Klasseneinteilungen aufgrund der Normalisierung ausgeglichen und so Klassifikationsergebnisse vergleichbar werden.

Ein solcher Vergleich von Klassifikatoren (Tan & Steinbach & Kumar 2006: 188ff.) ist in dieser Arbeit neben der Feststellung von Feature-Importance-Werten für ein Feature-Ranking ein Instrument zur Überprüfung, welches textstrukturelle Modell welche Apriori-Kategorisierung der Texte (funktional, situativ, diskursiv) besser vorhersagt. **Kappa** wird hier also zum Vergleich der Diskriminationsstärke verschiedener Feature-Sets bzgl. der Textkategorisierungen eingesetzt (vgl. Lapshinova-Koltunski & Zampieri 2018; Grzybek & Kelih & Stadlober 2005), allerdings in einer Version für den allgemeinen, nicht-binären Fall (Fleiss' Kappa, s. Fleiss 1971; vgl. Reiter & Frank & Hellwig 2014: 596). Zur Einordnung der Kappa-Werte dient folgende Interpretation bei Landis & Koch (1977: 165; vgl. auch Carletta 1996: 252; Sim & Wright 2005: 264):

- 0.0–0.2 Kappa: *slight agreement*
- 0.2–0.4 Kappa: *fair agreement*
- 0.4–0.6 Kappa: *moderate agreement*

15 S. Carletta 1996: 252: „The kappa coefficient (K) measures pairwise agreement among a set of coders making category judgments, correcting for expected chance agreement: $K = \frac{P(A)-P(E)}{1-P(E)}$ where $P(A)$ is the proportion of times that the coders agree and $P(E)$ is the proportion of times that we would expect them to agree by chance [...]. When there is no agreement other than that which would be expected by chance, K is zero. When there is total agreement, K is one."

- 0.6–0.8 Kappa: *substantial agreement*
- 0.8–1.0 Kappa: *almost perfect agreement*

4.3.4 Visualisierung und Diskriminanzanalysen

4.3.4.1 Klassen-gruppierte Average-Scores-Barplots

Wie in 4.2.4.2 ausgeführt, können Average-Scores-Barplots u. a. auch zur gruppierten Darstellung von Feature-Durchschnittswerten gemäß verschiedener Apriori-Genre-Klassifizierungen verwendet werden.

4.3.4.2 Lineare Diskriminanzanalyse als Visualisierungsmethode

Mit der **Linearen Diskriminanzanalyse** (LDA) können – ähnlich wie mit der Hauptkomponentenanalyse im *unsupervised*-Fall des Clusterings – Feature-Sets in Abhängigkeit von Klassen analysiert werden, indem die lineare Trennbarkeit der Daten in Abhängigkeit von der Kategorisierung modelliert wird. Dazu wird eine lineare Funktion (Diskriminanzfunktion) gesucht, d. h. die lineare Kombination der Merkmale, die die abhängige Variable (die Klasse) am besten vorhersagt (Beyerer & Richter & Nagel 2017: 173ff.; Schütze & Hull & Pedersen 1995: 231).[16]

Die lineare Diskriminanzanalyse (basierend auf Fisher 1936) ist verwandt mit Methoden der Regressionsanalyse wie der logistischen Regression, die lineare Zusammenhänge zwischen kontinuierlichen abhängigen und unabhängigen Variablen untersucht. Neben dem hier vorgestellten Einsatz als **Dimensionsreduktionsmethode** (Feature-Projection) kann LDA auch als Klassifikator eingesetzt werden (vgl. Vorgehen bei Grzybek & Kelih & Stadlober 2005). Die LDA kann dabei zur Klassifizierung mit mehreren Klassen eingesetzt werden und eignet sich auch bei wenigen Datensätzen (James u. a. 2017: 138).

Wie die Hauptkomponentenanalyse im unüberwachten Fall ermöglicht die Lineare Diskriminanzanalyse im überwachten Fall eine Dimensionsreduktion, nämlich eine Projektion in einen 1- oder 2-dimensionalen Raum – bei der LDA in den durch die erste oder die ersten beiden linearen Diskriminanten aufgespannten Raum. Sie eignet sich also auch zur Visualisierung von Feature-Sets in Abhängigkeit von den Klassen, auf die diese Feature-Sets durch die im Rahmen der Klassifikation berechneten LDA-Funktion abgebildet werden, und man kann so die Trennung durch den Klassifikator visuell beurteilen.

16 In der Auswertung in Kapitel 6 wurde aus Platzgründen auf die Verwendung der LDA-Methode verzichtet; sie wird an dieser Stelle aber aus Gründen der Vollständigkeit kurz erläutert (auf eine detaillierte Darstellung der formalen mathematischen Grundlagen wird entsprechend verzichtet; s. dazu etwa James u. a. 2017: 138ff.).

4.4 Sequenzrepräsentation und -klassifizierung

Neben der Repräsentation textstruktureller quantitativer Eigenschaften von Texten in Feature-Sets über das Datenrepräsentationsmodell (4.1.1), in dem von der linearen Abfolgestruktur der Texte abstrahiert wird, können Texte auch bzgl. ihrer sequentiellen Struktur über **sequenzbasierte** Repräsentations- und Klassifizierungsverfahren analysiert werden (vgl. Gabadinho u. a. 2011; Aggarwal 2015: 501ff., 521ff.). Sequenzen sind allgemein Folgen von kategorialen oder numerischen Werten; im Fall textstruktureller Sequenzrepräsentationen sind dies (vgl. 3.1.2) insbesondere textweite (globale) **kategoriale Folgen** von *tags* entsprechender Annotationskategorien (z. B. Folgen von Verbklasse-Labels als Ereignisabfolgen) bzw. **Partitur-Folgen** als numerische Folgen der Frequenzen einer linguistischen Einheit bzgl. ihres Auftretens in einer übergeordneten Einheit im Text (z. B. Folgen der Frequenzen von Clauses in direkter Rede pro Satz als Muster textinterner Diskursstrukturierung). Neben solchen globalen Wertefolgen können über eine Extraktion sog. **Frequent-Patterns** auch Modelle lokaler Sequenzinformationen erstellt werden, in denen häufig in den Texten auftretende Teilsequenzen bzw. Teilstrings textstruktureller Annotationsdaten als Merkmale in einem Feature-Set zusammengefasst werden.

Im Gegensatz zu den Feature-basierten Textstrukturrepräsentationen ermöglichen diese Sequenzrepräsentationen die Analyse von syntagmatisch-positionellen Strukturinformationen. Insbesondere basieren Frequent-Pattern-Muster auf Frequenzinformationen lokaler sequentieller Kookkurrenzen, während Partitur-Folgen eine Mischung aus Frequenz-akkumulierten Informationen und Informationen kodieren, die den globalen Textverlauf als sequentielle Anordnung von Äußerungen betreffen. Dabei finden hier – je nach sequentiellem Datentyp sowie dem Skalenniveau der Werte der Sequenz – unterschiedliche Distanzmaße und Klassifizierungsmethoden Anwendung.

4.4.1 Extraktion von Sequenzen

4.4.1.1 Extraktion von kategorialen Sequenzen

Im Fall einer Repräsentation der Textstruktur über kategoriale ***tag*-Sequenzen** wird – anstelle einer Transformation in Feature-Sets – die im Text gegebene, diskrete Folge der Labels der entsprechenden Annotationsgröße aus dem Korpus extrahiert und in einer Listenstruktur als geordnete Menge von Elementen abgespeichert:

```
NONE-NONE-SPEECH-NONE-SPEECH-SPEECH-NONE-SPEECH-NONE
```

Auflistung 4.1: Kategoriale Sequenzrepräsentation von Clauses in direkter Rede (Beispiel)

Im Gegensatz zur Textrepräsentation über Feature-Sets, bei der die Textlängennormierung im Rahmen der Feature-Construction geschieht (s. 4.1.3.1), findet eine Nor-

mierung bei dieser Sequenzrepräsentation erst während des Prozesses der Berechnung der Distanzmatrix statt.

4.4.1.2 Extraktion von numerischen Sequenzen

Analog zu den kategorialen Folgen werden **Partitur-Folgen** (Sequenzen von Frequenzen einer linguistischen Einheiten) als Listen diskreter numerischer Werte formalisiert:

```
0-0-1-0-1-1-0-1-0
```

Auflistung 4.2: Numerische Sequenzrepräsentation von Clauses in direkter Rede (Beispiel)

Auf diesen numerischen Folgen unterschiedlicher Längen kann dann durch die Anwendung des aus der Zeitreihen-Analyse stammenden Algorithmus des **Dynamic-Time-Warping** (DTW) eine Längennormierung durchgeführt werden; aus der dabei vorgenommenen Alignierung leitet sich gleichzeitig ein Distanzmaß ab, das für Clustering und Klassifikation verwendet werden kann (für Details s. 4.4.2.2 und 6.1.3.3).

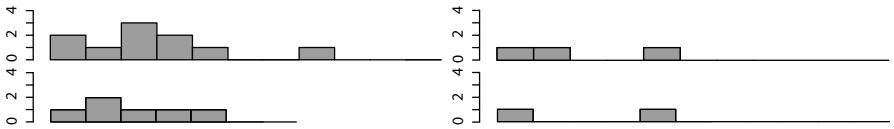


Abbildung 4.10: Beispiel ähnlicher Partitur-Folgen unterschiedlicher Länge

Durch geeignete Transformation von kategorialen Sequenzen in numerische Wertfolgen (u. a. durch Kodierung mit Dummy-Variablen, s. Abschnitt 6.1; vgl. James u. a. 2017: 129f.) kann die DTW-Normierung auch auf textstrukturelle Wertverlauf-Repräsentationen angewendet werden, um eine Distanzmatrix mit paarweisen Abständen zwischen den ursprünglich kategorialen Wertfolgen zu berechnen. Vor der Berechnung der DTW-Distanzen können die numerischen Wertefolgen für ein besseres Ergebnis der explorativen Datenanalyse durch den Ausschluss von Ausreißern auch mit geeigneten **Smoothing**-Methoden geglättet werden (Tukey 1977; vgl. Thompson 2011: 175ff.; Brillinger 2011: 535).

4.4.2 Clusteringmethoden für Sequenzen

Wie bei der Clusteranalyse von Feature-Sets wird auch bei der Clusteranalyse von Sequenzen (vgl. Gabadinho u. a. 2011: 32, Aggarwal 2015: 501) eine auf diesen berechnete Distanzmatrix, also eine Sammlung der paarweisen Distanzen zwischen den Sequenzen, benötigt. Anders als bei den Feature-basierten Repräsentationen, bei denen der Abstand zwischen den durch die Merkmalsvektoren definierten Datenpunkten im Merkmalsraum Grundlage der Distanzberechnung ist, müssen

bei sequentiellen Textstrukturrepräsentationen allgemein Unterschiede (Distanzen) zwischen kategorialen bzw. numerischen Wertfolgen berechnet werden.

4.4.2.1 Clustering kategorialer Sequenzen mit Edit-Distance-Maßen

Als Operationalisierung für den Abstand zwischen textweiten (globalen) kategorialen Folgen kommen insbesondere **Edit-Distance**-Maße der Stringanalyse in Frage (s. Aggarwal 2015: 501f.), die definiert sind als „the cost of edit operations required to transform one sequence into another" (Aggarwal 2015: 501; vgl. auch Deza & Deza 2016: 215ff.).[17] Die Editierdistanz ist also bestimmt über die Anzahl der Editieroperationen, die nötig sind, um eine diskrete kategoriale Folge (einen String) in eine andere zu transformieren (s. Gabadinho u. a. 2011: 25). Zu den Editieroperationen gehören u. a. *indel* (*insertion* oder *deletion*), *replacement*/*substitution*, *swap* und *move* (s. Deza & Deza 2016: 215f.). Je nach Operation können ggf. Kosten definiert werden, also eine Gewichtung der Operation als Maß für den Aufwand, eine Sequenz in eine andere zu transformieren (vgl. auch 6.1.3.2).

Klassische Edit-Distance-Maße sind die Levenstein- und die Hamming-Distanz (Deza & Deza 2016: 216, 225). In Kapitel 6 findet ein anderer Vertreter der Editierdistanzmaße (s. Aggarwal 2015: 82f.) Anwendung, nämlich die **Optimal-Matching**-Distanz (OM, Abbott & Tsay 2000). Das OM-Distanzmaß erlaubt durch die Verwendung der *indel*-Operation (*insertion* oder *deletion*) den Vergleich von Sequenzen unterschiedlicher Länge (Gabadinho u. a. 2011: 27; für Details s. 6.1.3.2). Dieses Edit-Distance-Maß stammt ursprünglich aus dem Bereich der Genetik (*Sequence-Alignment* mit Algorithmus von Needleman & Wunsch 1970) und wurde später auf Anwendungen in den Sozialwissenschaften übertragen (Abbott & Tsay 2000).

Das Vorgehen zur Berechnung der Clustergruppen, ausgehend von einer OM-Distanzmatrix zwischen Sequenzen, geschieht dann grundsätzlich mit den oben beschriebenen Methoden des hierarchischen Clusterings; analog zum Vorgehen bei Gabadinho u. a. (2011: 32) wird allerdings in Kapitel 6 für das Clustering kategorialer Sequenzdaten Wards Agglomerationsmethode verwendet. Für die Berechnung der Distanzmatrix werden in diesen globalen Sequenzanalysen zunächst paarweise die Optimal-Matching-Abstände zwischen Sequenzen berechnet und anschließend diese berechneten Abstände normiert (vgl. 4.1.3.1); für die OM-Distanzen wird in Kapitel 6 eine Normierung durch die Länge der längeren Sequenz (*maxlength*) gewählt (vgl. Gabadinho u. a. 2011: 29; Abbott & Tsay 2000; s. auch 6.1.3.2). Während bei Feature-Sets also bereits textlängennormierte Feature-Werte zur Distanzberechnung zwischen den Textrepräsentationen verwendet werden, geschieht die Normalisierung hier erst nach der Berechnung der Distanzen.

17 Dagegen können bei den in 4.4.4 folgenden lokalen Sequenzanalysen über Frequent-Patterns die bereits vorgestellten Distanzmaße für frequenzbasierte Features verwendet werden (vgl. Aggarwal 2015: 503f.).

4.4.2.2 Clustering mit Dynamic-Time-Warping-Distanz

Dynamic-Time-Warping (DTW) ist eine u. a. in der Verarbeitung von Sprachsignalen zentrale Methode der Längennormierung für Zeitreihen, bei der numerische Wertfolgen unterschiedlicher Länge gestaucht und aufeinander abgebildet werden (s. Xing & Pei & Keogh 2010: 42; Beyerer & Richter & Nagel 2017: 38; Aggarwal 2015: 79ff.):

> Dynamic time warping (DTW) is a sequence alignment method allowing a nonlinear mapping of one sequence to another by minimizing the (total cumulative) distance between them. Used originally in speech recognition, DTW is applied now to temporal sequences of video, audio, and graphics data. (Deza & Deza 2016: 408)

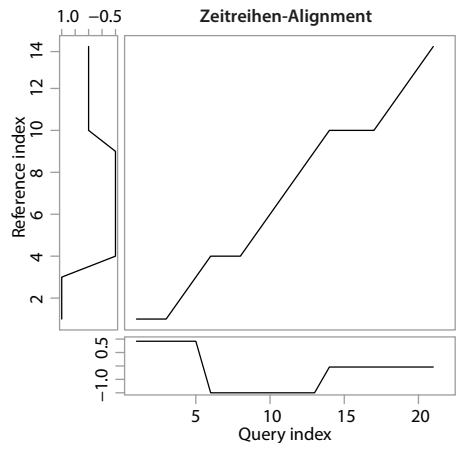


Abbildung 4.11: Alignment von zwei Zeitreihen-Textrepräsentationen mit Dynamic-Time-Warping

Bei der Anwendung des Dynamic-Time-Warping-Algorithmus (Sakoe & Chiba 1978) zur Abbildung numerischer Folgen aufeinander wird gleichzeitig auch ein Maß für die Distanz zwischen den Folgen berechnet, also für den Aufwand der Transformation der einen in die andere Sequenz (s. Giorgino 2009: 3). Dieses Distanzmaß kann dann für ein **Sequenz-Clustering** (Aggarwal 2015: 501f.) eingesetzt werden, d. h. „[...] the number of *nonmatches* between the two sequences can be used with dynamic time warping. [...] The idea is to stretch and shrink the time dimension dynamically to account for the varying speeds of data generation for different series“ (Aggarwal 2015: 501, Hervorhebung im Original).

Durch paarweise Berechnungen der DTW-Distanzen zwischen numerischen Sequenzrepräsentationen kann also deren Distanzmatrix erstellt werden; auf diese DTW-Distanzmatrix können dann die oben vorgestellten hierarchischen Clusteringmethoden angewendet werden (s. auch 6.1.3.3).

4.4.3 Klassifikationsmethoden für Sequenzen

4.4.3.1 Distanzbasierte Sequenzklassifikation mit k-Nearest-Neighbour

Für die Klassifikation von Sequenzen (s. Aggarwal 2015: 522) kann der k-Nearest-Neighbour-Klassifikator (**knn**) als distanzbasierte Klassifikationsmethode (s. Aggar-

wal 2015: 522; Xing & Pei & Keogh 2010: 42) verwendet werden.[18] Dazu wird dieser Nächste-Nachbarn-Algorithmus auf eine Distanzmatrix (hier: DTW- oder OM-Distanzen von textstrukturellen Sequenzen, vgl. Ding u. a. 2008: 1546) angewendet.

K-Nearest-Neighbour ist eine einfache, parameterfreie Klassifikationsmethode – also ohne Annahme einer Wahrscheinlichkeitsverteilung –, die zur Vorhersage der Klassen für ein neues Objekt die k gemäß der Distanzmatrix nächstgelegenen Objekte auswählt und das neue Objekt der Klasse zuordnet, der die Mehrheit dieser k Nächsten-Nachbarn angehören; bei 1-Nearest-Neighbour wird das Objekt einfach der Klasse zugeordnet, der das nächstgelegene Objekt angehört (s. James u. a. 2017: 39ff.; Beyerer & Richter & Nagel 2017: 154ff.; Manning & Schütze 1999: 604; Manning & Raghavan & Schütze 2009: 297ff.).

4.4.3.2 Sequenzklassifikation mit Spectrum-String-SVM

Eine weitere Methode zur Klassifikation von sequentiellen Daten ist die Verwendung von **Spectrum-String-SVMs** (Leslie & Eskin & Noble 2002; s. auch Hofmann & Schölkopf & Smola 2008: 12f.; Aggarwal 2015: 502, 524). Ein SVM-Klassifikator (Support Vector Machine) nimmt eine Klassifikation von Feature-repräsentierten Objekten vor, indem er eine Diskriminationsfunktion sucht, die im Feature-Space den Abstand zwischen einer zwischen den Datenpunkten verschiedener Klassen eingezogenen, diese trennenden Hyperebene und den nächstgelegenen Datenpunkten maximiert (Beyerer & Richter & Nagel 2017: 194); im zweidimensionalen Fall ist die Hyperebene eine Gerade. Diese nächstgelegenen Datenpunkte (Vektoren) bestimmen dabei die gesuchte Hyperebene (Beyerer & Richter & Nagel 2017: 195), man nennt diese auch *Stützvektoren*, daher der Name dieser Methode; vgl. Manning & Raghavan & Schütze 2009: 319: „An SVM is a kind of large-margin classifier: it is a vector space based Machine Learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data [...]."

Durch den sog. **Kernel-Trick**, d. h. durch Transformation in einen höherdimensionalen Feature-Raum, in dem man eine lineare Trennung finden kann (s. Hofmann & Schölkopf & Smola 2008; Beyerer & Richter & Nagel 2017: 199f.), kann die SVM-Methode als grundsätzlich lineare Methode (Hyperebene = lineare Funktion) für eine Anwendung auf nicht linear trennbare Daten erweitert werden und so eine lineare Trennbarkeit der Daten erzwungen werden. Dieser Kernel-Trick kann auch auf **Sequenzdaten** angewendet werden, indem man einen Feature-Raum konstruiert, dessen Dimensionen aus der Gesamtmenge von Teilsequenzen zu gegebener Länge k (= **Spektrum**) aus dem Alphabet der Sequenz bestehen (s. Xing & Pei & Keogh 2010:

18 knn könnte also auch auf die Feature-Set-basierten Repräsentationen angewendet werden, indem die für das Clustering erzeugte Distanzmatrix verwendet wird; allerdings wäre hier die Feature-Importance nicht direkt aus dem Modell extrahierbar wie bei Random-Forest als einer Embedded-Feature-Selection.

42; Hofmann & Schölkopf & Smola 2008: 11f.); diese Konstruktion basiert auf einer entsprechenden Kernel-Funktion, dem sog. **String-Kernel**: „Using appropriate kernel functions, an SVM classifier can even be used to classify complicated objects like genome sequences or the words of a (natural) language" (Beyerer & Richter & Nagel 2017: 206). Im Gegensatz zur distanzbasierten knn-Sequenzklassifikation ist diese Variante der Sequenzklassifikation also eine Art der Feature-basierten Klassifikation (s. Xing & Pei & Keogh 2010: 49f.) in einem dafür konstruierten hochdimensionalen Merkmalsraum von Teilsequenzen.

4.4.4 Frequent-Pattern-Extraktion und -Klassifizierung

Die zwei bisher vorgestellten Typen textstruktureller Repräsentationen sind zum einen das Datenrepräsentationsmodell, dessen Feature-Sets (insbesondere *bag*-Modelle) auf quantitativen Merkmalen linguistischer Einheiten basieren und entsprechend von der linearen Anordnung im Text abstrahieren; zum anderen sequentielle Modelle, die aus globalen, textweiten Folgen kategorialer oder numerischer Daten bestehen.

Durch die Operationalisierung textstruktureller Parameter über numerische Partitur-Folgen können auch frequenzbezogene Eigenschaften linguistischer oder informationsstruktureller Einheiten in ihrer linearen Anordnung, also in einem globalen sequentiellen Modell, erfasst werden. Umgekehrt bieten **Frequent-Pattern**-Modelle (s. Gabadinho u. a. 2011: 3; Aggarwal 2015: 503) die Möglichkeit, **lokale** sequentielle Abfolgemuster über deren Textfrequenzen als Merkmale eines Feature-Sets zu verwenden:

> The idea is that the frequent subsequences represent the key structural characteristics that are common across different sequences. After the frequent subsequences have been determined, the original sequences can be transformed into this new feature space, and a "bag-of-words" representation can be created in terms of these new features. (Aggarwal 2015: 503)

Für so ein „bag of frequent subsequences" (Aggarwal 2015: 503) können dann die Feature-basierten Klassifizierungsmethoden des hierarchischen Clusterings (als „subsequence-based clustering", Aggarwal 2015: 503) sowie der Random-Forest-Klassifikation (als Feature-basierte Sequenzklassifikation, s. Xing & Pei & Keogh 2010: 41f.) angewendet werden, insbesondere sind hier also Feature-Importance-Analysen der lokalen Sequenzmuster möglich. Außerdem können Frequent-Patterns auch mit nicht-sequentiellen Features in einem Feature-Set kombiniert werden (vgl. 6.6.1). Auch gegenüber globalen Sequenz-Modellen, deren Alignierungsmethodik gerade bei längeren Sequenz-Objekten (bzw. mit stark unterschiedlichen Längen) an Grenzen stößt, haben Frequent-Pattern-Modelle mit lokalen Teilsequenzen Vorteile, da

sie die Objekte aufgrund lokaler sequentieller Ähnlichkeiten abgleichen und so von irrelevanten Teilsequenzen absehen können:

> The major problem with the aforementioned methods is that they are based on similarity functions that use *global* alignment between the sequences. For longer sequences, global alignment becomes increasingly ineffective because of the noise effects of computing similarity between pairs of long sequences. Many local portions of sequences are noisy and irrelevant to similarity computations even when large portions of two sequences are similar. (Aggarwal 2015: 503, Hervorhebung im Original)

Für die Extraktion und Analyse von sequentiellen Teilmustern werden Methoden des Frequent-Sequential-Pattern-Minings verwendet (s. Aggarwal 2015: 493ff.; Tan & Steinbach & Kumar 2006: 429ff.; Han u. a. 2007; Mabroukeh & Ezeife 2010; Mooney & Roddick 2013).[19] Zentral ist dabei die Auswahl von **häufigen** Teilmustern von Sequenzobjekten, d. h. solcher lokalen Teilsequenzen, deren Vorkommen in den Objekten über einem festgelegten Schwellenwert liegt (s. Aggarwal 2015: 495; Mabroukeh & Ezeife 2010: 5; vgl. 6.1.2.6): „Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold" (Han u. a. 2007: 56). Diese Mindesthäufigkeit wird auch als **Support** bzw. Minimal Support bezeichnet; meist werden noch weitere Constraints für die Extraktion von Frequent-Patterns verwendet, z. B. eine Maximallänge:

> Determine the frequent subsequences F from the sequence database D using any frequent sequential pattern mining algorithm. Different applications may vary in the specific constraints imposed on the sequences, such as a minimum or maximum length of the determined sequences. (Aggarwal 2015: 503)

In der Fallstudie in Kapitel 6 werden im Rahmen der Extraktion häufiger Ereignisabfolgen aus einer globalen Textsequenz von Verbklasse-*tag*-Folgen (z. B. `M-M-A-A--S-A-A`) zusammenhängende Teilfolgen (N-Gramme, vgl. Aggarwal 2015: 503) von **Übergängen** (*transitions*, vgl. Gabadinho u. a. 2009: 16) von einem Zustand in einen anderen extrahiert (`(M>A)-(A>S)-(S>A)`). Damit wird eine zusätzliche Reduktion einer Folge von Ereignistypen auf den Wechsel (Übergang) zwischen den verschiede-

19 Der Begriff Frequent-Pattern wird unterschiedlich weit gefasst; Aggarwal 2015 bezieht ihn primär auf das Itemset-Mining in Assoziationsanalysen und behandelt die Methoden des hier relevanten Sequential-Pattern-Minings in einem separaten Kapitel (Aggarwal 2015: 493ff.: „Mining Discrete Sequences"); er weist aber auf die Ähnlichkeit beider Verfahren hin (2015: 498f.). Bei Han u. a. 2007 werden Frequent-Patterns dagegen weiter gefasst und beinhalten auch Teilsequenzen, s. Han u. a. 2007: 56: „Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern."

nen Typen vorgenommen, um Blöcke (Chunks, vgl. 3.7.3) der Übergänge zwischen Ereignistypen zu erhalten, die unabhängig von der Verweildauer in einem Zustand (Ereignistyp) sind (für Details s. 6.1.2.6). Report 4.4.1 zeigt das Ergebnis der Extraktion solcher häufiger Ereignisübergänge aus dem obugrischen Korpus.

	Support	Count	subseq
1	0.88	30.00	(ACT>MOTION)
2	0.85	29.00	(MOTION>ACT)
3	0.74	25.00	(SPEECH>ACT)
4	0.71	24.00	(ACT>MOTION)-(MOTION>ACT)
5	0.68	23.00	(MOTION>ACT)-(ACT>MOTION)

Report 4.4.1: Extraktion häufiger Ereignisübergänge (Beispiel)

Üblicherweise wird abschließend eine Feature-Subset-Selection angewendet (s. Xing & Pei & Keogh 2010: 41; vgl. 4.3.3) bzw. wird die Anzahl der Frequent-Patterns bereits durch Constraints bei der Auswahl beschränkt, vgl. Aggarwal 2015: 503: „Typically, a subset of frequent subsequences should be selected, so as to maximize coverage and minimize redundancy. The idea is to use only a modest number of relevant features for clustering." In dieser Arbeit werden für die Extraktion von häufigen Ereignisabfolgen nur die Übergangsfolgen extrahiert, die einen sehr hohen Support im Korpus (0.67, s. 6.1.2.6) haben, also in den meisten Texten des Korpus vorkommen (mindestens in 2/3 aller Texte), um mit diesen häufigsten Ereignisabfolgen diejenigen Strukturmuster herauszufiltern, die für die narrativen Kerntexte des obugrischen Korpus *typisch* sind (vgl. 5.2.3). Die Anzahl der Features wird hier also bereits durch dieses Pruning, d. h. das Abschneiden von Features unter einem bestimmten Schwellenwert, reduziert (vgl. Guyon & Elisseeff 2003: 1158). Bei der Verwendung der Frequenzdaten der Frequent-Patterns als Feature-Werte müssen auch die unterschiedlichen Textlängen berücksichtigt werden; durch Verwendung einer Kodierung der bloßen An- bzw. Abwesenheit des Merkmals (Presence/Absence) wie in Kapitel 6 kann diese Textlängennormierung entfallen (s. 6.1.2.6).

4.4.5 Visualisierung und Analyse von Sequenzdistributionen

Für die Visualisierung der Clusteringergebnisse kategorialer globaler Sequenzanalysen sowie zur Darstellung von nach Textsortenkategorisierungen geordneten Sequenzen kommen spezifische Visualisierungsmethoden für kategoriale Sequenzdistributionen zum Einsatz: das sind vor allem gruppierte **Sequenz-Indexplots** (s. Plot 6.7.4 für ein Beispiel) sowie Plots der Durchschnittszeiten in den verschiedenen Zuständen (vgl. Gabadinho u. a. 2011: 12ff.). Für DTW-basierte Clusteranalysen numerischer Folgen kann durch Berechnung von **Barycenter-Kurven** (s. Petitjean & Ketterlin & Gançarski 2011) der durchschnittliche Kurvenverlauf der Partiturrepräsentationen pro Clustergruppe dargestellt werden (s. Plot 6.7.3 für ein Beispiel).

5 Daten und Annotationsmethoden

Kapitelzusammenfassung

Dieses Kapitel gibt eine Übersicht über die Korpusdaten, die in der in Kapitel 6 folgenden Fallstudie für die Erprobung der Methoden und Parameter für eine Text-Weltmodell-bezogene Genre-Mustererkennung verwendet werden. Dazu gehört eine Kurzeinführung in die soziolinguistische Situation sowie eine Übersicht über das typologische Profil der beiden obugrischen Sprachen Khanty und Mansi, aus denen die Texte des Erprobungskorpus stammen. Es folgt eine Diskussion sowohl der Textdatenauswahl dieses in der Arbeit verwendeten Korpus als auch der verschiedenen, in der explorativen Korpusanalyse in Kapitel 6 als Vergleichsdaten verwendeten Apriori-Genre-Einteilungen dieser Texte. Das Kapitel wird abgeschlossen durch eine Übersicht über die Korpusannotationen, auf denen die Berechnung der textstrukturellen Parameter in Kapitel 6 basiert.

5.1 Khanty und Mansi

5.1.1 Soziolinguistische Situation

Die beiden zur uralischen Sprachfamilie gehörenden obugrischen Sprachen **Khanty** (auch: Chantisch oder Ostyak) und **Mansi** (auch: Vogul) werden noch von einigen tausend Sprechern in Nordwestsibirien an den Flussläufen und Nebenflüssen von Ob und Irtyš östlich des Urals gesprochen (s. Abbildung 5.2). Das Verbreitungsgebiet (s. Abbildung 5.1) umfasst primär den Autonomen Kreis der Chanten und Mansen in der Russischen Föderation (historisch auch: Jugra) sowie das Gebiet nördlich davon an der Mündung des Ob (Abondolo 1998b: 358; Keresztes 1998: 389). Die obugrischen Sprachen sind stark bedroht, einige Dialekte sind bereits verschwunden. In den letzten drei Jahrhunderten waren die Khanten und Mansen einem starken kulturellen und sprachlichen Einfluss des Russischen ausgesetzt, entsprechend herrscht Bilingualismus vor (s. Kolga u. a. 2020a; 2020b).

Abbildung 5.1: Die obugrischen Sprachen und ihre Dialektgebiete im Einzugsgebiet des Obs, Nordwestsibirien (reproduziert nach Abondolo 1998a: xxvii, Karte iii)

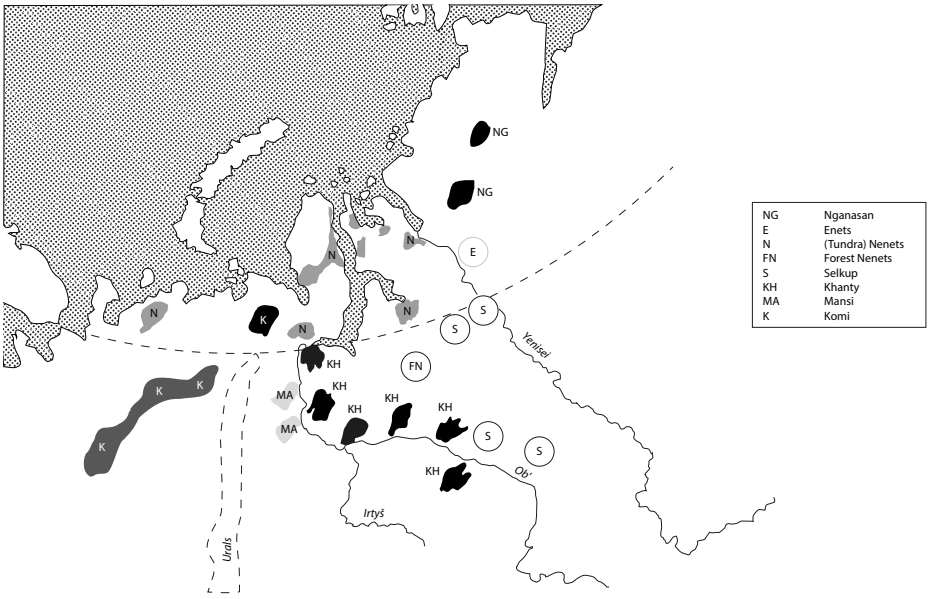


Abbildung 5.2: Khanty (KH) und Mansi (MA) im Kontext der uralischen Sprachen Nordwestsibiriens (reproduziert nach Abondolo 1998a: xxviii, Karte iv)

Innerhalb der uralischen Sprachfamilie kann man mit Abondolo 1998a die ugrische Subgruppe als deren Kern verorten: „Hungarian, Mansi, and Khanty are the sole survivors of what is here seen as the core, i. e. most central and innovating region, of

Uralic linguistic and cultural space" (Abondolo 1998a: 6). Die obugrischen Sprachen und die obugrische Kultur haben sich dabei anders entwickelt als das Ungarische, das sich früh von diesem **ugrischen Kern** abgespalten hatte und durch räumliche Trennung (Wanderung aus Westsibirien nach Süden und dann Westeuropa) und durch intensiven Kontakt mit verschiedenen, insbesondere westeuropäischen Sprachgruppen geprägt und beeinflusst wurde:

> As vehicles of culture, both Hungarian and proto-ObUgrian suffered major blows in the form of radical restructuring of *genre de vie*: while speakers of proto-ObUgrian, in consequence of their migration east and north, were thrust back into a neolithic cultural frame, speakers of Hungarian underwent the reverse scenario, namely the accelerated modernization which attended their settling in central Europe. The effect on the shared lexicon has been catastrophic: in both cases, old discourse was replaced or transmuted, usually beyond recognition […]. (Abondolo 1998a: 6, Hervorhebung im Original)

Obwohl auch die sprachhistorische Trennung der beiden obugrischen Sprachen relativ lange zurückliegt (ungefähr im 13. Jh., s. Kolga u. a. 2020a), zeigen sie durch ihren Verbleib in Westsibirien[1] und den durch diese Nachbarschaft bedingten kulturellen und sprachlichen Austausch (vgl. Csepregi 2009: 15) sowohl sprachlich als auch kulturell eine enge Verwandtschaft (s. Abondolo 1998a; Honti 1998; Keresztes 1998). Historisch bestand Sprachkontakt mit den umliegenden Sprechergemeinschaften (etwa den samojedischsprachigen Nenzen und Tungusisch-Sprechern im Osten sowie Komi-Sprechern im Westen) sowie im Süden mit Turk-Sprechern (Abondolo 1998b: 382f.; Keresztes 1998: 422f.; vgl. Skribnik 2010).[2]

Die Sprechergemeinschaften der Chanten und Mansen teilen also eine ähnliche **Lebenswelt** (vgl. Hatto 2017a); dazu gehört u. a. die Subsistenz durch Jagd und Fischfang (einige Gruppen auch durch Viehhaltung, insbesondere Rentierhaltung; s. Niko-

1 Vgl. allerdings Lintrop 2020: „The rich military vocabulary of the Ob-Ugrians and several motifs in their mythology and folklore imply contacts with the cattle-breeding cultures of the steppe area. Presumably, not all the Ugrian cattle-breeders made their way together with the Huns or after them through the steppes between the Urals and the Aral and Caspian Seas into Europe to evolve into Hungarian tribes in the grassy plains of Bashkiria, but part of them were forced by the flux of the Great Migration to the West Siberian Lowlands where they assimilated among the local Ugrian hunters and fishers."

2 Vgl. Lintrop 2020: „Later influences from Turkic peoples must be taken into account, too – in the 16th century, several Ob-Ugrian princes were forced to pay tribute to the Siberian khan and participate in the military ventures of the Khanate." Vgl. auch Keresztes 1998: 422: „There is evidence for Siberian Tatar influence on Southern dialects of Mansi from the fourteenth century onwards. There are more than 500 Tatar loans, embracing most social and economic semantic spheres […]. The main current of Tatar influence was broken by the Russian conquest of western Siberia in the sixteenth century."

laeva 1999: 3; Lintrop 2020),[3] eine schamanistisch-animistische Tradition (s. Chernetsov 1963; textuell ausgedrückt u. a. in rituellen Liedern; vgl. auch die Tier- und Zaubermärchen im Korpus), ein gemeinsames Mythen- und Tabusystem (vgl. dazu etwa die Texte „Why one shouldn't misbehave at night" im Korpus) sowie ein ausgeprägter Totemkult: „[…] to this day the Khanty are the custodians of the world's most elaborate bear cult" (Keresztes 1998: 424; vgl. Hatto 2017a).

Abbildung 5.3: Obugrische Lebenswelt: Der Große Jugan und Jurty Kajukovy (© Zs. Schön, 2015)

Abbildung 5.4: Obugrische Lebenswelt: Zubereiten eines Hechts, Derevnja Taurovo (© Zs. Schön, 2015)

Aufgrund dieser geteilten Lebenswelt kann man annehmen, dass die Texte der mansischen sowie der khantischen, in der Mündlichkeit verwurzelten **narrativen Tradition** in ähnlichen Sprachgebrauchssituationen produziert und rezipiert werden und damit auch eine ähnliche textstrukturelle Typisierung (d. h. ein ähnliches Text-Weltmodell) aufweisen, vgl. Schütz & Luckmann 2003: 318: „Jedermann ist in eine Situation hineingeboren, in der ihm die Sprache, genauer, eine bestimmte Sprache, als eine Komponente der historischen Sozialwelt vorgegeben ist." Da die narrative Sprachpraxis von Khanten und Mansen noch grundlegend in der mündlichen Überlieferung verankert ist (s. Hatto 2017a; Cushing 1980: 215),[4] also in der ursprünglichen Praxis des Geschichtenerzählens, eignen sich die obugrischen Volkserzählungen als

3 S. Nikolaeva 1999: 3: „The modern Ostyaks usually live in small villages, which sometimes consist of only a few houses. Their main occupations are fishing and hunting; the northernmost Ostyaks also engage in reindeer breeding." S. auch Lintrop 2020: „In the 16th and 17th centuries, the Mansi got their living mainly from hunting. Fishing played a less significant role in their economy than among the Khanty. An important field of activity was forest apiculture. The hive trees were private property; […] The Western and Southern Mansi bred cattle and tilled soil. The Northern Mansi also went in for reindeer herding."

4 Vgl. Cushing 1980: 215: „[…] in the Ugrian tradition the term 'literature' implies oral literature. Moreover the expected distinction between prose and verse has little meaning; instead there is a division into 'song' and 'tale', the former having musical accompaniment and the latter none." Für Untersuchungen zum kulturellen Hintergrund insbesondere der khantischen mündlichen Literaturtradition s. auch Siikala & Ulyashev 2011 und Hatto 2017b.

Datengrundlage für die Rekonstruktion der der Produktion dieser Texte zugrunde liegenden strukturellen Typisierungen des Sprachgebrauchs (vgl. Abschnitt 1.2), da sie näher an der ursprünglichen mündlichen Erzählpraxis[5] dieser Sprechergemeinschaft sind als etwa europäische Märchensammlungen und man annehmen kann, dass ein solches Volksmärchen „die entsprechende linguistische Praxis unmittelbar abbildet" (Schulze 2019: 16).

5.1.2 Sprachtypologische Kurzübersicht

Folgender Abschnitt gibt einen knappen sprachtypologischen Abriss der beiden obugrischen Sprachen Khanty und Mansi.[6] Vornehmlich werden hier die für die konkrete Operationalisierung textstruktureller TWM-Parameter in Kapitel 6 relevanten Sprachspezifika der grammatikalischen und informationsstrukturellen Kodierung des Obugrischen behandelt. Die Darstellung orientiert sich dabei an dem bei Skopeteas u. a. (2006) entwickelten Fragenkatalog zur Untersuchung von informationsstrukturellen Forschungsfragen.

Die obugrischen Sprachen sind **agglutinierend** und zeigen eine *head-final*-Konstituentenstruktur; Adverbiale werden entsprechend neben Lokalkasussuffixen auch über Postpositionen markiert (vgl. Schön 2017). In der Kodierung der primären syntaktischen Funktion zeigen Khanty und Mansi einen **akkusativischen** Alignment-Typ (mit dialektaler Ausnahme im Ostkhanty, vgl. Honti 1998: 351). Morphosyntaktische Mittel zur Markierung der grammatischen Relationen sind Wortstellung (SOV) und Agreement über Personalsuffixe (Person und Numerus; auch Dual), wobei hier zwei Konjugationstypen auftreten (**Agreement-Split**), nämlich monopersonales Subjekt-Agreement (sog. *subjektive* Konjugation) sowie polypersonales Subjekt-Objekt-Agreement (sog. *objektive* Konjugation; s. Honti 1998: 347f.). Eine Akkusativ-Kasusdifferenzierung von Subjekt und Objekt gibt es nur im pronominalen Bereich (mit dialektaler Ausnahme).

Die obugrischen Sprachen sind **pro-drop**-Sprachen, d. h. die Realisierung von Subjekt und Objekt, z. B. durch pronominalen Ersatz, ist nicht obligatorisch; es treten also regelmäßig **Nullanaphern** auf, insbesondere in Subjekt-Position. Die Notwendig-

5 Einige Dialekte sind zwar im 20. Jh. verschriftlicht worden, die Schrifttradition ist aber eher eingeschränkt, vgl. Virtanen & Sosa 2018: 237: „Different variants [*of Mansi and Khanty*] have different backgrounds of written use: some of them are regularly used in media and literature, and some have only been written by scholars." Vgl. ebenso Nikolaeva 1999: 3: „The Ostyaks began to use written records in the 1930s, based on the Cyrillic alphabet. However writing and reading skills in the native language are not widespread." Die hier verwendeten Daten haben aber (mit Ausnahme zweier Zeitungstexte, s. 5.2.2) als Aufzeichnungen aus Feldforschung (zum großen Teil von Beginn des 20. Jh.s) einen Hintergrund in einer spezifischen *oral tradition*.

6 Für Darstellungen der Grammatiken s. u. a. Honti 1998 (Übersicht der Grammatik beider obugrischer Sprachen); Abondolo 1998b und Nikolaeva 1999 (Khanty); sowie Keresztes 1998, Riese 2001 und Virtanen 2015 (Mansi).

keit der Realisierung des Objekts ist dabei abhängig vom Konjugationstyp[7] – es liegt also ein Fall differentieller Objekt-Markierung vor (**DOM**, Bossong 1998; vgl. Virtanen 2014: 391; Skribnik 2010). Die Verwendung des Konjugationstyps wiederum ist konditioniert durch den Diskursstatus der Referenten: Die objektive Konjugation ist obligatorisch, wenn das Objekt *secondary topic* ist (Nikolaeva 2001).[8]

Neben diesem Agreement-Split werden im Obugrischen auch Diathesen als informationsstrukturelle Kodierungsmittel verwendet; so besitzen Khanty und Mansi eine **Passiv-Diathese**, die regelmäßig zur Topikalisierung verwendet wird (Promotion Topik in Subjekt-Position; s. Honti 1998: 351f.; Skribnik 2010). Des Weiteren tritt ein **Dativ-Split** auf, wodurch die Promotion des Rezipient-Arguments einer Ditransitivkonstruktion (z. T. auch anderer topikaler Argumente; meist *secondary topics*) in Objekt-Position durch Demotion des Patiens-Arguments in den adverbialen Bereich ermöglicht wird (Skribnik 2010: 50; vgl. auch Skribnik 2001).

In der Nominalphrase treten als spezifizierende Angaben Adjektive (qualitative Spezifizierung) sowie **Possessivsuffixe** auf (referentielle Spezifizierung, vgl. Janda 2019); eine Genus- oder andere Klassenmarkierung gibt es nicht, auch ein Artikel im Sinne eines obligatorischen Determinierers existiert nicht, stattdessen wird Definitheit primär über Agreement-Splits kodiert (vgl. Skribnik 2010). Außerdem können Demonstrativpronomen sowie weitere Pronomen auch als Determinierer verwendet werden (und dienen dann ggf. als Fokusmarker). Die (nicht-obligatorischen) Personalpronomen werden u. a. zur (z. B. kontrastiven) Fokussierung eingesetzt (emphatische Pronomen statt Nullmarkierung wie für den Hauptreferenten).

Neben modalen Kategorien wird am Verb **Tempus** grammatikalisch markiert (s. Honti 1998: 346; Mansi hat Präsens- und Past-Markierung, Khanty hat Präsens-Markierung, Past ist dort nullmarkiert).

Subordination wird (ebenso wie Relativsätze) über nicht-finite verbale Konstruktionen, insbesondere Partizipien realisiert (Keresztes 1998: 420f.; Abondolo 1998b: 380f.). Selten gibt es (unter Einfluss des Russischen) auch subordinierte finite Sätze mit Subjunktion (vgl. Riese 2001: 70).

Zusammenfassend ist im Obugrischen als sprachspezifische Textstruktur-relevante Kodierungsstrategie (vgl. Schulze 2019: 30) insbesondere eine Kodierung der **Topikalität** bzw. allgemeiner des Informationsstatus der Referenten (d. h. „Definitheitsaspekte, die nicht nur topikale Ketten, sondern auch Bezüge auf das Weltwissen der Hörerschaft betreffen", Schulze 2019: 30) über morphosyntaktische Strategien wie Konjugations-Split, Diathesen oder z. T. auch Wortstellung (Fokusposition) statt

7 Bei objektiver Konjugation kann das Objekt entfallen, vgl. die beiden Nullanaphern für Subjekt und Objekt im letzten Clause von Interlinearversion 2 in 5.3.6.

8 Vgl. auch den Exkurs zur differentiellen Topikalitätsmarkierung anhand der Daten zur referentiellen Distanz in transitiven Sätzen des Obugrischen in 3.6.3, insbesondere Abbildung 3.4.

über Definitheitsmarker festzustellen (vgl. Skribnik 2001; Nikolaeva 2001; Virtanen 2014).

5.2 Korpusdaten und Auswahlkriterien

5.2.1 Quantitative Basisdaten

Datengrundlage für die in dieser Arbeit durchgeführte explorative Fallstudie sind 34 obugrische Texte, die im Rahmen des am Institut für Finnougristik der LMU München beheimateten Forschungsprojektes **OUDB** (Ob-Ugric Database)[9] unter Mitarbeit des Autors syntaktisch, semantisch und informationsstrukturell annotiert wurden (s. Korpusverzeichnis). Diese Auswahl der vollannotierten Texte aus dem OUDB-Gesamtkorpus umfasst knapp 11 000 Tokens bzw. 8 000 Worttokens (durchschnittliche Textlänge 37 Sätze, Median 25 Sätze).

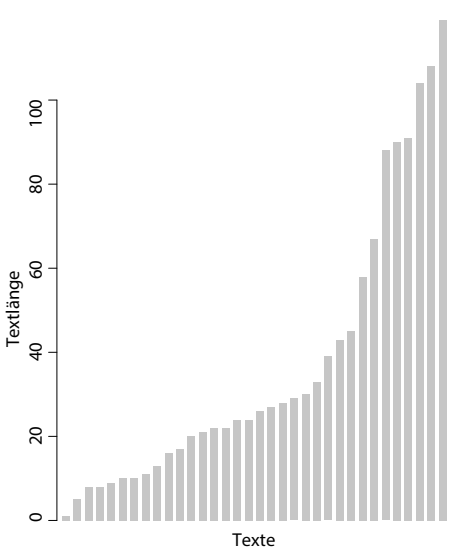


Abbildung 5.5: Textlängen im Korpus (in Sätzen)

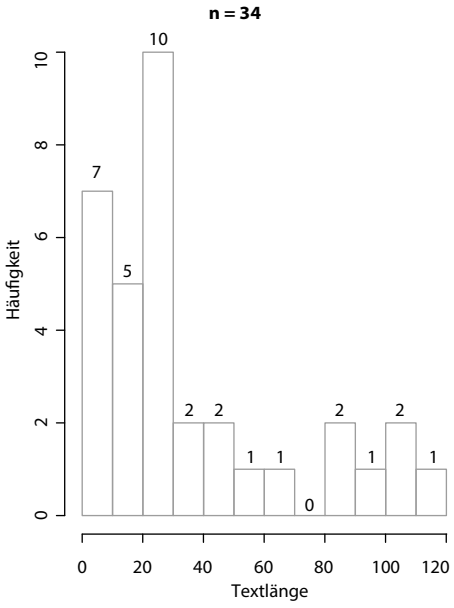


Abbildung 5.6: Histogramm der Textlängen im Korpus (in Sätzen)

9 Website des OUDB-Forschungsprojekts unter http://www.oudb.gwi.uni-muenchen.de.

Text-ID	Genre-Form	Dialekt	Titel-Code	Sätze	Worttoken	Token
728	**eth**	SK	**Hunting**	16	135	168
730	**eth**	SK	**Cold**	30	241	318
732	tal	SK	LittleBird	39	273	374
741	**myt**	NM	**Creation**	1	11	15
742	**myt**	NM	**Fireflood**	22	203	267
750	**myt**	NM	**SosvaRaid**	91	817	1113
889	tal	NM	Southcountry	88	519	658
1076	**eth**	SK	**MakeBread**	26	270	327
1231	**jou**	NM	**Faraway**	43	435	511
1232	tal	NM	BeardedMan	17	105	132
1233	tal	NM	Woodpecker	45	249	342
1237	tal	NM	ThreeSons	119	621	830
1262	tal	PM	Bullfinch	104	616	889
1263	tal	NV	FourSisters	108	532	770
1314	tal	PA	LittleBird	90	419	578
1346	tal	SK	BeardedMan	8	56	95
1347	tal	SK	Cranberry	5	29	52
1348	**jou**	SK	**Guest**	28	208	259
1352	tal	YK	TwoFires (TAK)	11	97	132
1355	**eth**	YK	**Trap**	9	71	94
1368	**fas**	PM	**GameVoicedSong**	10	34	45
1373	**fas**	PM	**HazyDaySong**	13	59	71
1459	tal	YK	TwoFires (AIK)	24	191	264
1462	tal	YK	TwoFires (TMJ)	22	146	182
1469	tal	YK	TwoFires (JFP)	33	285	385
1483	tal	YK	Cranberry (OAL)	20	108	156
1484	tal	YK	Cranberry (AIK)	10	53	75
1488	tal	YK	Cranberry (AJM)	8	70	99
1493	tal (mixed)	YK	Cranberry (JFP)	21	118	181
1514	eth *(tal)*	YK	OldDogMan (TMK)	58	323	520
1515	eth *(tal)*	YK	OldDogWoman (AIK)	27	146	190
1518	eth *(tal)*	YK	OldDogWoman (SPK)	67	349	481
1523	tal	YK	Misbehave (AJM)	24	130	175
1542	tal (mixed)	YK	Misbehave (JFP)	29	164	236
Gesamt:				1 267	8 083	10984

Tabelle 5.1: Übersicht der quantitativen Basisdaten des Korpus (Abweichungen von der Genre-Einteilung in dieser Arbeit unterstrichen; Texte nicht-narrativer Genres in Fettdruck)

5.2.2 Kontextuelle Metadaten

Die khantischen Texte im Korpus stammen aus zwei der östlichen Khanty-Dialekte (vgl. Abbildung 5.1): Zum einen aus dem **Surgut-Khanty** – das sind Texte aus der Feldforschung von 1993–2011 (Sammlung CS) sowie ein journalistischer Text von 2012 (Sammlung XY) –, zum anderen aus dem **Yugan-Khanty** – hier sind es

Texte aus der Feldforschung von 2010–2017 (Sammlungen AZ und ZS) und 1901 (Sammlung PA).[10]

Die mansischen Texte des Korpus stammen zum einen aus den beiden (inzwischen ausgestorbenen) westlichen Mansi-Dialekten des **Pelym-Mansi** – das sind Texte aus der Feldforschung von 1888–1889 (Sammlung MU) und 1901–1906 (Sammlung KL) – sowie des **North-Vagilsk-Mansi**, ebenfalls mit Texten aus der Feldforschung der Sammlung KL von 1901–1906. Zum anderen sind Texte aus dem **Nordmansischen** vertreten, auch hier solche aus der Feldforschung von 1901–1906 (Sammlung KL) und ab 1923 (Sammlung CH) sowie ein journalistischer Text von 2005 (Sammlung LS).[11]

Das Korpus bietet damit einen knappen Querschnitt durch die verschiedenen obugrischen Sprachen und ihre Dialektgruppen; so sind etwa Dialekte aller bei Virtanen & Sosa 2018 bzgl. ihrer informationsstrukturellen Kodierungstypik unterschiedenen arealen Gruppen vertreten.

Mit Ausnahme zweier Zeitungstexte (`jou`)[12] stammen die Textdaten sowohl aus historischer als auch aktueller Feldforschung. Die Texte können größtenteils dem Genre der **Volkserzählungen** zugeordnet werden; dies sind also narrative Texte aus mündlicher Erzähltradition. Darüber hinaus sind im Korpus auch einige ethnographische Berichte (`eth`), mythologische Erzählungen (`myt`) sowie sog. Fate Songs (`fas`) vertreten. Unterschiede in der Transkription aufgrund von unterschiedlichen Feldforschungsstandards, etwa bzgl. der Phonematisierung, sind im OUDB-Korpus durch ein dialekt- und sammlungsübergreifendes Transkriptionskonzept vereinheitlicht.[13]

Code	Sammlung	Jahr(e)	Dialekt	Sprache	Genres
MU	Munkácsi, Bernát	1888–1889	PM	Mansi	fas
KL	Kannisto & Liimola	1901–1906	NM, PM, NV	Mansi	tal, myt
CH	Chernetsov, Valeri	ab 1923	NM	Mansi	magic_tal/anim_tal
LS	Zeitung *Luima Seripos*	2005	NM	Mansi	jou
PA	Paasonen, Heikki	1901	PA (YK-1901)	Khanty	magic_tal
CS	Csepregi, Márta	1993–2011	SK	Khanty	magic_tal/anim_tal, eth
AZ	Kayukova & Schön	2010–2017	YK	Khanty	magic_tal/anim_tal
ZS	Schön, Zsófia	2010–2016	YK	Khanty	magic_tal/anim_tal, eth
XY	Zeitung *Khanty Yasang*	2012	SK	Khanty	jou

Tabelle 5.2: Übersicht der im Korpus vertretenen Sammlungen

10 Khanty hat bzw. hatte drei Hauptdialektgruppen (Nord, Süd und Ost; die südliche ist bereits verschwunden).

11 Mansi hat bzw. hatte vier Hauptdialektgruppen (Nord, Süd, Ost und West; die westliche und die südliche sind bereits verschwunden).

12 Vgl. Genre-Kürzel in Tabelle 5.2.

13 Für die Annotation und Analyse der grammatischen und informationsstrukturellen Größen ist die Transkription ohnehin sekundär.

Tabelle 5.3 gibt eine Übersicht über die Herkunft (Dialekt, Sammlung, Jahr) und den Titel in englischer Übersetzung (eine deutsche Übersetzung liegt nicht für jeden Text vor). Die in dieser Arbeit verwendeten Titel-Codes dienen vor allem dem Einsatz in der Clustering-Visualisierung zur schnellen Identifizierung von Varianten; die vollständigen Angaben inkl. Verlinkung zum Text im Onlinekorpus sind im Korpusverzeichnis zu finden.

5.2.3 Textsorten des Korpus

Den Kern des Erprobungskorpus machen obugrische **Volkserzählungen** unterschiedlicher dialektaler, zeitlicher und räumlicher Herkunft aus; häufig liegen diese in mehreren Varianten vor.[14] Eine Verzerrung der TWM-Typik durch den persönlichen Stil der Sprecher (vgl. Schulze 2004a: 557; Schwarz-Friesel & Consten 2014: 23, 141ff.) sollte bei Volkserzählungen als mündlichen Überlieferungen weniger relevant sein; auch ein editorischer Einfluss sollte bei den hier in der Feldforschung aufgezeichneten, nicht bzw. nur schwach redigierten Texten im Hintergrund stehen. Solche möglichen diastratischen Einflüsse sollen aber im Rahmen der Auswertung mit berücksichtigt werden.

Festzustellen ist, dass sich die im Korpus vertretenen Varianten eines Märchens – die neben Sprechervarianten häufig gleichzeitig auch diachrone und sprachliche bzw. dialektale Varianten sind – untereinander relativ stark unterscheiden (sichtbar schon an der häufig sehr unterschiedlichen Textlänge). Entsprechend ist hier von einer Nacherzählung des Plots der Geschichte auszugehen (vgl. Abschnitt 6.7) statt von einer Wiedergabe des exakten Texts aus dem Gedächtnis (vgl. Panzer 2020: § 30; Schulze 2004b: 206).[15] Als solche „spontaneous reformulations of narrative traditions" (Schulze 2004b: 206) kodieren diese Varianten – trotz der sprachlichen und textuellen Abweichungen – dasselbe (bzw. ein sehr ähnliches) tradiertes kognitives Text-Modell; die Text-Varianten einer solchen *story* sollten entsprechend eine ähnliche schematische Struktur aufweisen (d. h. eine ähnliche *story grammar*, vgl. Rumelhart 1975). Entsprechend sollten sie auch in den schematischen Strukturparametern des kognitiven Text-Weltmodells (TWM) übereinstimmen, das (nach Annahme hier) die Textproduktion in der Reproduktion des erinnerten Plots als spezifische Handlungsfolgenstruktur leitet (vgl. Propp 1972: 27, 112).[16]

14 Die einzelnen Varianten werden im Detail in Abschnitt 6.7 besprochen; identifizierbar sind die Varianten eines Märchens über den entsprechenden Titel-Code (z. T. mit Zusatzangabe des Sprechers), vgl. etwa die `Cranberry`-Varianten in Tabelle 5.1.

15 Vgl. Schulze 2004b: 206: „the tale […] is not 'entrenched' as such, but remembered for its plot."

16 In dieser Arbeit werden textstrukturelle TWM-Eigenschaften untersucht, keine substantiellen wie Einleitungs- oder andere Formeln im Sinne von „ritualized passages" (Schulze 2004b: 206; vgl. 1.1.3); zu solchen substantiellen Märchenmarkern im Obugrischen vgl. Sauer 2004–2005 („Formeln und Formelhaftes in ostjakischen Märchen").

Text-ID	Genre-Form	Dialekt	Sprache	Sammlung	Jahr	Titel-Code	Titel
728	**eth**	SK	Khanty	CS	1996	**Hunting**	Hunting Adventure
730	**eth**	SK	Khanty	CS	1996	**Cold**	An Old Story
732	tal	SK	Khanty	CS	1992	LittleBird	The Little Bird and His Sister (SK)
741	**myt**	NM	Mansi	KL	(1901–1906)	**Creation**	Creation of the Earth
742	**myt**	NM	Mansi	KL	(1901–1906)	**Fireflood**	The Holy Fireflood
750	**myt**	NM	Mansi	KL	(1901–1906)	**SosvaRaid**	The Middle Sosva Old Man's Raid to the Sacred Site on the Water
889	tal	NM	Mansi	CH	(ab 1923)	Southcountry	Southcountry
1076	**eth**	SK	Khanty	CS	(1993–2011)	**MakeBread**	Make Bread
1231	**jou**	NM	Mansi	LS	(2005)	**Faraway**	We Studied in a Faraway Country
1232	tal	NM	Mansi	CH	(ab 1923)	BeardedMan	A Tale of Four Men
1233	tal	NM	Mansi	CH	(ab 1923)	Woodpecker	The Poor Old Woman, Her Husband and the Woodpecker
1237	tal	NM	Mansi	CH	(ab 1923)	ThreeSons	In Olden Times There Lived a Man Who Had Three Sons
1262	tal	PM	Mansi	KL	(1901–1906)	Bullfinch	There Was an Old Man and an Old Woman
1263	tal	NV	Mansi	KL	(1901–1906)	FourSisters	Four Sisters and a Man with His Daugther
1314	tal	PA	Khanty	PA	1901	LittleBird	The Little Bird and His Sister (PA)
1346	tal	SK	Khanty	CS	1992	BeardedMan	Two Short Stories 1
1347	tal	SK	Khanty	CS	1992	Cranberry	Two Short Stories 2
1348	**jou**	SK	Khanty	XY	(2012)	**Guest**	Our Guest
1352	tal	YK	Khanty	AZ	2010	TwoFires(TAK)	Two Domestic Fires (TAK)
1355	**eth**	YK	Khanty	ZS	2010	**Trap**	Trap
1368	**fas**	PM	Mansi	MU	(1888–1889)	**GameVoicedSong**	Song Composed by Agsinia Apanasejevna
1373	**fas**	PM	Mansi	MU	(1888–1889)	**HazyDaySong**	Song Composed by Agafia Stefanovna
1459	tal	YK	Khanty	AZ	2015	TwoFires(AIK)	Two Domestic Fires (AIK)
1462	tal	YK	Khanty	AZ	2015	TwoFires(TMJ	Two Domestic Fires (TMJ)
1469	tal	YK	Khanty	AZ	2015	TwoFires(JFP)	Two Domestic Fires
1483	tal	YK	Khanty	AZ	2012	Cranberry(OAL)	Little Cranberry (OAL)
1484	tal	YK	Khanty	ZS	2015	Cranberry(AIK)	Little Cranberry (AIK)
1488	tal	YK	Khanty	AZ	2015	Cranberry(AJM)	Little Cranberry (AJM)
1493	tal (mixed)	YK	Khanty	AZ	2015	Cranberry(JFP)	Little Cranberry (JFP)
1514	eth *(tal)*	YK	Khanty	ZS	2012	OldDogMan(TMK)	Old-Dog-Backtendoned-Man (TMK)
1515	eth *(tal)*	YK	Khanty	AZ	2015	OldDogWoman(AIK)	Old-Dog-Backtendoned-Woman (AIK)
1518	eth *(tal)*	YK	Khanty	AZ	2015	OldDogWoman(SPK)	Old-Dog-Backtendoned-Woman (SPK)
1523	tal	YK	Khanty	AZ	2015	Misbehave(AJM)	Why One Shouldn't Misbehave at Night (AJM)
1542	tal (mixed)	YK	Khanty	AZ	2015	Misbehave(JFP)	Why One Shouldn't Misbehave at Night (JFP)

Tabelle 5.3: Übersicht der Metadaten des Korpus (Abweichungen von der Genre-Einteilung in dieser Arbeit unterstrichen; Texte nicht-narrativer Genres in Fettdruck)

Die Volkserzählungen im Korpus lassen sich in zwei Subgenres (vgl. Schulze 2020: 592) unterteilen: Eine Gruppe von **Zaubermärchen** (vgl. ATU-Gruppe 300–749),[17] die im Sinne Propps (1972) einen spezifischen Handlungsstrukturtyp mit „anecdotic character […] [*which*] corresponds to the standard scheme that has the positive protagonist being involved in a 'dramatic' event from which (s)he escapes under guidance of a helper" (Schulze 2004a: 555) darstellen; sowie eine Gruppe von kurzen, fabelartigen **Tiermärchen** mit z. T. belehrendem Charakter (vgl. ATU-Gruppe 1–299),[18] die sich auszeichnen durch einen für die Fabel typischen antithetischen Aufbau sowie durch eine auf die wesentlichen Handlungen der Gegenspieler verdichtete Handlungsstruktur ohne Exposition (Dithmar 1974: 100, 191). Inwiefern sich diese erzählstrukturellen Subtypen in der TWM-Typisierung widerspiegeln, wird sich in den Feature-Analysen mittels der zuvor beschriebenen Clustering- und Klassifikationsmethoden in Kapitel 6 zeigen.

Neben diesen Volksmärchen sind einige Texte angrenzender Textsorten, also folkloristische Texte im weiteren Sinne vertreten (s. Nguyen u. a. 2012: 379; vgl. Bascom 1965),[19] nämlich die **mythologischen** Sagen und legendenartigen Märchen (`myt`, vgl. ATU-Gruppe 750–849) sowie die lyrischen, autobiographischen **Fate Songs** (`fas`).[20] Außerdem sind einige **ethnographische** Texte (`eth`) als mündliche persönliche Berichte und Erzählungen sowie zwei in obugrischen Zeitungen veröffentlichte **Reiseberichte** (`jou`) als nicht-mündliche Vergleichsdaten enthalten. Insgesamt dienen die Texte der verschiedenen peripheren Genres im Rahmen der explorativen Textstruktur-Analyse dieser Arbeit als Kontrollgruppe für die Volksmärchen, deren textstrukturelle TWM-Typik es im Rahmen dieser Arbeit mit den in Kapitel 4 vorgestellten Datenexplorationsmethoden insbesondere zu untersuchen gilt.

In dieser Arbeit werden zwei Textsorten-Kategorisierungen angesetzt: zum einen die auf der Auszeichnung der Textsorten im OUDB-Korpus basierende Einteilung `BASE` (Tabelle 5.4),[21] zum anderen die ebenfalls auf dieser Einteilung basierende

17 ATU = Aarne-Thompson-Uther-Index (Uther 2011) als Motiv-basierte Typisierung von Märchentexten.

18 S. auch Panzer 2020: 27 und Lüthi 1998: 24 zur Differenzierung narrativer folkloristischer Texte.

19 Vgl. Bascom 1965: 3 (Hervorhebung im Original): „*Prose narrative*, I propose, is an appropriate term for the widespread and important category of verbal art which includes myths, legends, and folktales. […] Prose narrative is clearly less equivocal for this broad category than 'folktale' because the latter has so often been used by folklorists to mean *Märchen*."

20 Diese „Personal Songs" (Ojamaa & Ross 2004: 134) sind „a poetic genre found among the Ob-Ugrians, in which men and women tell the story of their lives" (Birnbaum 1977: 231). Vgl. auch Ojamaa & Ross 2004: 134: „In 1901–1906, a Finnish linguist and folklorist Juha Artturi Kannisto organised expeditions to visit the Mansis. He denoted autobiographical songs as fate songs. Kannisto characterizes these as lyric songs, where the moods, experiences and fate of the author (a man or a woman) are described. He adds that these songs are also called vodka-drinking songs […]."

21 S. Spalte „Genre-Form" in Tabelle 5.3; allerdings werden davon abweichend die Texte 1514, 1515 und 1518 (drei Varianten eines Textes über das Ungeheuer Amp-Chun-Lonep) aufgrund ihrer Funktionstypik als fiktive Erzählungen als Märchen (`tal`) klassifiziert – genauer als Zaubermärchen (`magic_tal`) – anstatt als persönliche Erzählungen (`eth`). Text 1514 wird auch explizit vom Sprecher als Märchen benannt.

Kategorisierung `GENRE` (Tabelle 5.5), die aber durch Unterscheidung der beiden Subgenres Zaubermärchen (`magic_tal`) und Tiermärchen (`anim_tal`) innerhalb der Klasse der Erzähltexte (`tal`) eine narratologische Subdifferenzierung vornimmt.

Ethnographische Texte (`eth`)
• mündliche persönliche Berichte oder Erzählungen • nicht-fiktiv
Personal Songs / Fate Songs (`fas`)
• autobiographische Lieder, mündlicher Vortrag • lyrisch
Journalistische Berichte (`jou`)
• schriftliche persönliche Berichte • nicht-fiktiv
Mythologische Sagen (`myt`)
• mündlich überlieferte, mit realen Orten verknüpfte Handlung (vgl. Panzer 2020: § 21; Bascom 1965: 4) • fiktive Historie
Volksmärchen (`tal`)
• mündlich überlieferte Erzählung mit phantastischer, raum-zeitlich unbestimmter Handlung (vgl. Panzer 2020: § 1; Bascom 1965: 4) • fiktive Handlung

Tabelle 5.4: Klassen der BASE-Kategorisierung

Fabel-Tiermärchen (`anim_tal`)
• Tiere, Pflanzen oder Dinge als stereotype Charaktere (vgl. Dithmar 1974: 110f.) • fabelähnliche Handlung • kurze Zeitspanne, ein Handlungsort (vgl. Lüthi 1998: 68; Dithmar 1974: 103f.; Uther 2011)
Zaubermärchen (`magic_tal`)
• meist Menschen und übersinnliche Wesen als Pro- bzw. Antagonisten • längere Handlungsketten, Wechsel zwischen verschiedenen Orten (vgl. Propp 1972; Uther 2011)

Tabelle 5.5: Narrative Subklassen der GENRE-Kategorisierung

5.2.4 Diskursfunktionale Apriori-Kategorisierungen

Im Rahmen der in dieser Arbeit beabsichtigten Analyse von kognitiven Textstrukturtypen wurden die Texte des Erprobungskorpus vor der eigentlichen Feature-Analyse zusätzlich zu der Einteilung nach literarischen Textsorten auch gemäß textlinguistisch begründeter, diskursfunktionaler Genre-Kategorisierungen manuell gelabelt (vgl. 1.1.2 und 2.1.3). Mit Hilfe dieser Metadaten unterschiedlicher textfunktionaler

Genre-Einteilungen der Texte des Korpus kann einerseits untersucht werden, wie gut bestimmte, aus den Korpusdaten gewonnene Klassifikationsmodelle der kognitiven Texttypologie-Parameter jeweils diese verschiedenen, auf Grundlage literaturwissenschaftlicher bzw. textlinguistischer Kriterien postulierten Texttypen vorhersagen (s. Abschnitt 4.3).[22] Darüber hinaus können diese theoretisch begründeten Genre-Einteilungen auch im Rahmen einer externen Clusterevaluation (s. 4.2.3) eingesetzt werden, um den Grad der Übereinstimmung zwischen diesen verschiedenen **Apriori-Typologien** mit anhand von TWM-Parametern in den Daten als induktive Texttypologie gefundenen Cluster-Gruppen zu bestimmen.

Die Auswahl dieser Apriori-Genre-Kategorisierungen für das Korpus orientiert sich dabei einerseits an den bereits vorgestellten literaturwissenschaftlichen Textsorten-Einteilungen, die man mit Heinemann & Viehweger (1991: 170) als eine ursprünglich alltagstheoretisch verankerte Funktionstypisierung von Texten verstehen kann, in der inhaltliche sowie pragmatisch-funktionale Merkmale vermischt sind.[23] Theoretisch reflektiert, basieren solche Textsorten-Einteilungen auf elementaren Typen der Textfunktion (etwa die Informationsfunktion, die Appellfunktion oder die ästhetische Textfunktion; s. Brinker 2000: 176; Bußmann 2008: 721) mit entsprechenden Haupttextsorten wie Nachrichtentext, Gebrauchstext oder Literatur; subdifferenziert nach textinternen Merkmalen inhaltlich-formaler Art ergeben sich dann die klassischen Textsorten: Erzählung, Drama, persönlicher Bericht, Interview, Reportage, Kochrezept usw. (s. Bußmann 2008: 727; Schwarz-Friesel & Consten 2014: 40).

Als eine weitere, über textfunktionale Kategorien begründete Texttypologie werden sog. **Textstrategietypen** (s. Brinker 2000: 183f.) einer funktionalen Text- und Diskurslinguistik herangezogen, die auf textinternen, diskursstrukturellen Merkmalen basieren und als deren Grundtypen üblicherweise die Narration, die Deskription und die Argumentation angesetzt werden.[24] Als Grundlage der entsprechenden

22 Vgl. auch das Vorgehen bei Grzybek & Kelih & Stadlober 2005 sowie auch die Diskussion zur Einteilung nach verschiedenen Klassifikationssystemen bei Kessler & Nunberg & Schütze 1997: 34.

23 Vgl. Heinemann & Viehweger 1991: 170: Textsorten als „Durchschnittserfahrungen (von Sprechern einer bestimmten Kommunikationsgemeinschaft) […] als globale sprachliche Muster zur Bewältigung von spezifischen kommunikativen Aufgaben in bestimmten Situationen […][,] globale […] Textstrukturmuster"; eine solche Definition von Textsorten ähnelt der in dieser Arbeit angenommenen, gebrauchsbasierten Bestimmung von induktiv gewonnenen TWM-Texttypen als Genres (im Gegensatz zu Texttypen als theoretischen Typkonstrukten); vgl. dazu auch Fix 2008: 131ff.: Textsorten als „Organisationsformen des Alltagswissens"; vgl. auch Schwarz-Friesel & Consten 2014: 44 zur Differenz von Textsorte und Texttyp.

24 S. de Beaugrande & Dressler 1981 und Heinemann & Viehweger 1991; in solchen Textstrategie-Theorien, die Texte auffassen „als Anweisungen dafür, wie ein Text als Struktur zu verstehen ist, wie ein Wirklichkeitsmodell hergestellt werden kann" (Hartung 2000: 84), sind insbesondere informationsstrukturelle Muster des durch den Text repräsentierten kognitiven Modells relevant (vgl. Schwarz-Friesel & Consten 2014: 42), so z. B. die thematische Entfaltung (vgl. Bußmann 2008: 730). Solche Textmusterkriterien (Fix 2008: 67f.) ermöglichen also durch Rekonstruktion der Textfunktion über eine entsprechende Strukturanalyse (Brinker 2000: 183f.) die Differenzierung verschiedener Textstrategien (vgl. Bußmann 2008: 735f.; Schwarz-Friesel & Consten 2014: 145).

DISC_STRUCT-Kategorisierung dieser Arbeit dient die auf diskursstrukturellen Kriterien basierende Texttypologie von Longacre (1983), der eine Typisierung des „overall purpose of the discourse“ (1983: 3) anstrebt.

Diskurstyp	Abfolge	Orientierung
narr	chronologisch	Partizipant
expos	logisch	Thema
behav	logisch	Partizipant
proc	chronologisch	Thema (Ziel)

Tabelle 5.6: Diskurstypen-Einteilung nach Longacre 1983

Die Einteilung von Longacre (s. Tabelle 5.6) basiert dabei auf den zwei binären informationsstrukturellen Kriterien der Abfolge (chronologische vs. logische Ereignisstruktur)[25] sowie der Orientierung (Partizipanten- vs. Thema-orientierte Referentenstruktur),[26] wodurch sich die vier Texttypen **narrativ** (narr), **expositorisch** (expos), **verhaltensbezogen** (behav) sowie **prozedural** (proc) ergeben, denen die Texte des Korpus anhand der binären Kategorien zugeordnet werden (s. Tabelle 5.7).

Narrativer Diskurs (narr)
• Partizipanten-orientiert • chronologische Abfolge • die meisten narrativen Texte (tal) sowie einige eth-Texte und ein myt-Text
Expositorischer Diskurs (expos)
• Thema-orientiert • logische Abfolge • zwei myt-Texte und die fas-Songs
Verhaltensbezogener Diskurs (behav)
• Partizipanten-orientiert • logische Abfolge • die jou-Texte sowie einige eth- und tal-Texte (in Narration eingebettete Verhaltensnormen)
Prozeduraler Diskurs (proc)
• Thema-orientiert • chronologische Abfolge • einige eth-Text (z. B. *Making Bread*)

Tabelle 5.7: Klassen der DISC_STRUCT-Kategorisierung

25 Vgl. Longacre 1983: 3: „Contingent temporal succession (henceforth contingent succession), refers to a framework of temporal succession in which some (often most) of the events or doings are contingent on previous events or doings.“

26 Vgl. Longacre 1983: 3: „Agent orientation refers to orientation towards agents […], with at least a partial identity of agent reference running through the discourse.“

Diese Kategorisierung von Longacre korrespondiert mit den u. a. bei de Beaugrande & Dressler (1981: 190f.) sowie Heinemann & Viehweger (1991: 237) angesetzten Textstrategie-Grundtypen: die **Narration** als chronologische Kette von Ereignissen (Handlungen von Referenten); die **Deskription** als räumlich-logische Beziehungssetzung von Objekten und Sachverhalten, ähnlich dem expositorischen Diskurs bei Longacre 1983 (vgl. auch Werlich 1975: 30f., 35f.; Heinemann 2000: 360);[27] sowie die **Argumentation** als logische Beziehungssetzung von Ereignissen und Handlungen, ähnlich dem verhaltensbezogenen Diskurs bei Longacre: „Behavioral discourse (a broad category including exhortation, eulogy and political speeches of candidates) is minus in regard to contingent succession but plus in regard to agent orientation (it deals with how people did or should behave)" (Longacre 1983: 3). In anderen Texttypologien werden als weitere Textstrategie-Typen die **Explikation** (Brinker 2018: 69ff.) bzw. die **Instruktion** (Werlich 1975: 70f.) unterschieden, die ungefähr dem prozeduralen Diskurs im Sinne einer kausal-zeitlichen Abfolge bzgl. Zielen und Objekten bei Longacre entsprechen: „Procedural discourse (how to do it, how it was done, how it takes place) is plus in respect to contingent succession (the steps of a procedure are ordered) but minus in respect to the agent orientation (attention is on what is done or made, not on who does it)" (Longacre 1983: 3).

Als zweite diskursfunktionale Apriori-Typologie wird eine textexterne Kategorisierung herangezogen, die auf die äußeren Umstände der Kommunikationssituation bezogen ist, in der ein Text als Abfolge von Sprachhandlungen stattfindet. Mögliche Kriterien einer solchen kommunikationstheoretisch begründeten Typologie sind die Art des Kommunikationsmediums oder der Typ des Adressatenkreises; in dieser Arbeit wird für die textexterne Diskurskategorisierung `COMM_SIT` eine binäre **Registereinteilung** (persönlich vs. öffentlich) bzgl. des Bekanntheitsgrads und der Anzahl der Kommunikationspartner gewählt (vgl. Bußmann 2008: 730).[28] Die Zuordnung der Texte zu einer dieser binären Kategorien ist nicht immer selbstverständlich, sodass auch hier versucht wurde, der Auswahl bestimmte Kriterien zugrunde zu legen (s. Abbildung 6.4); dennoch sind die Übergänge zwischen den beiden Klassen unscharf, insbesondere bei den narrativen, mündlich vorgetragenen Texten (diese Problematik von Apriori-Typologien unterstreicht gleichzeitig auch die Wichtigkeit der induktiven Typologie-Erstellung durch explorative Clustermethoden als zentrale Methodik dieser Arbeit).

27 Werlich setzt dagegen die Exposition als eigenen, auf die „in Elemente zerlegende oder analytische Textstrukturierung" (1975: 36) bezogenen Texttyp an.

28 Vgl. Grzybek & Kelih & Stadlober 2005, die im Rahmen der Auswertung ihrer Korpusstudie auf die grundlegende Bedeutung einer solchen binären Registereinteilung für eine textstrukturelle Typologie hinweisen: „Es verstärkt sich damit die Annahme, daß die Trennung von öffentlichen/offiziellen vs. Privat- bzw. Alltagstil von hoher Relevanz ist." (Grzybek & Kelih & Stadlober 2005: 113)

Privat / Interview (`priv`)
• dazu gehören persönliche Berichte (`eth`) sowie narrative Texte aus der aktuellen Feldforschung (`YK`-Korpus) • diese narrativen Texte sind phonetisch getreue Transkriptionen der mündlichen Äußerungen auf Grundlage von Audioaufzeichnungen (enthalten u. a. auch die Selbstkorrekturen durch den Sprecher sowie den Dialog mit dem Interviewer: häufig stückweise Nacherzählung der Geschichte für den Interviewer auf Nachfrage)
Öffentlicher Vortrag (`publ`)
• intendiert für öffentlichen Vortrag (bzw. für Veröffentlichung editiert) • dazu gehören journalistische Texte (`jou`), Lieder (`fas`) und mythologische Texte (`myt`) sowie Erzähltexte aus älteren Sammlungen

Tabelle 5.8: Klassen der COMM_SIT-Kategorisierung

Wie oben ausgeführt, dienen die Apriori-Kategorisierungen in dieser Arbeit als Analyse- und Vergleichsgrößen einerseits dem Abgleich der konkurrierenden, theoretisch begründeten Textgruppierungen mit den im Clustering rein induktiv in den Daten festgestellten Gruppierungen; andererseits können diese Apriori-Kategorisierungen durch Verwendung in Klassifikationsmodellen auch untereinander dahingehend verglichen werden, wie gut verschiedene Feature-Sets jeweils diese konkurrierenden Textsorteneinteilungen unterscheiden und welche Merkmale für eine funktionale Apriori-Typisierung jeweils am wichtigsten sind. Durch die Verwendung sowohl textinterner als auch textexterner funktionaler Apriori-Texttypologien wird dabei zum einen durch die textinterne, informationsstrukturbezogene Kategorisierung (`DISC_STRUCT`) eine Überprüfung dahingehend ermöglicht, ob sich deren inhaltlich-strukturelle Makrostrukturtypen (also etwa Textmuster der Topikkontinuität, der thematischen Progression usw.) auch entsprechend im Rahmen der automatischen Modellkonstruktion gemäß der linguistischen Annotationsdaten rekonstruieren lassen. Außerdem kann die automatische Analyse der Feature-Sets anhand des textexternen, auf den Situationstyp bezogenen Kriteriums (`COMM_SIT`) auch dazu dienen, die in dieser Arbeit getroffenen theoretischen Annahmen zum Lernen von TWM als Genre-Modelle durch kognitive Prozesse der Typisierung von Kommunikationssituationen datengestützt zu prüfen.

Text-ID	BASE	GENRE	DISC_STRUCT	COMM_SIT	Dialekt	Title-Code
728	**eth**	eth	narr	priv	SK	**Hunting**
730	**eth**	eth	proc	priv	SK	**Cold**
732	tal	magic_tal	narr	publ	SK	LittleBird
741	**myt**	myt	expos	publ	NM	**Creation**
742	**myt**	myt	expos	publ	NM	**Fireflood**
750	**myt**	myt	narr	publ	NM	**SosvaRaid**
889	tal	magic_tal	narr	publ	NM	Southcountry
1076	**eth**	eth	proc	priv	SK	**MakeBread**
1231	**jou**	jou	behav	publ	NM	**Faraway**
1232	tal	anim_tal	narr	publ	NM	BeardedMan
1233	tal	magic_tal	narr	publ	NM	Woodpecker
1237	tal	magic_tal	narr	publ	NM	ThreeSons
1262	tal	magic_tal	narr	publ	PM	Bullfinch
1263	tal	magic_tal	narr	publ	NV	FourSisters
1314	tal	magic_tal	narr	publ	PA	LittleBird
1346	tal	anim_tal	narr	publ	SK	BeardedMan
1347	tal	anim_tal	narr	publ	SK	Cranberry
1348	**jou**	jou	behav	publ	SK	**Guest**
1352	tal	anim_tal	narr	priv	YK	TwoFires (TAK)
1355	**eth**	eth	behav	priv	YK	**Trap**
1368	**fas**	fas	expos	publ	PM	**GameVoicedSong**
1373	**fas**	fas	expos	publ	PM	**HazyDaySong**
1459	tal	anim_tal	narr	priv	YK	TwoFires (AIK)
1462	tal	anim_tal	narr	priv	YK	TwoFires (TMJ)
1469	tal	anim_tal	narr	priv	YK	TwoFires (JFP)
1483	tal	anim_tal	narr	priv	YK	Cranberry (OAL)
1484	tal	anim_tal	narr	priv	YK	Cranberry (AIK)
1488	tal	anim_tal	narr	priv	YK	Cranberry (AJM)
1493	tal	anim_tal	narr	priv	YK	Cranberry (JFP)
1514	tal	magic_tal	behav	priv	YK	OldDogMan (TMK)
1515	tal	magic_tal	behav	priv	YK	OldDogWoman (AIK)
1518	tal	magic_tal	behav	priv	YK	OldDogWoman (SPK)
1523	tal	magic_tal	behav	priv	YK	Misbehave (AJM)
1542	tal	magic_tal	behav	priv	YK	Misbehave (JFP)

Tabelle 5.9: Übersicht der Textsorten-Einteilung des Korpus gemäß der vier Kategorisierungen BASE, GENRE, DISC_STRUCT und COMM_SIT (Abweichungen von BASE bzgl. der Genre-Form-Einteilung des Korpus unterstrichen; Texte nicht-narrativer Genres in Fettdruck)

5.3 Annotationsparameter und -methoden

Die Berechnung und Feature-Construction der kognitiven texttypologischen Parameter basiert auf basalen Annotationsparametern der morphologischen, syntaktischen, semantischen und pragmatisch-informationsstrukturellen Mikro-Ebene. Grammatikalische Informationen zu Morphologie und Syntax spielen vor allem eine Rolle für die Berechnung der globalen Parameter, die sich auf den Strukturaufbau der Texte durch linguistische Einheiten beziehen. Semantische und informationsstrukturelle Grunddaten sind Grundlage für die Berechnungen in der Feature-Construction der funktional motivierten Parameter, die sich auf die referentiell- und relationssemantischen sowie informationsstrukturellen Eigenschaften linguistischer Ausdrücke beziehen.

Die Annotation dieser Parameter im hier verwendeten obugrischen Korpus wurde im Rahmen des OUDB-Projekts mit zumeist halbautomatischen Methoden durchgeführt, für Details s. Wisiorek & Schön 2017 und Janda & Wisiorek & Eckmann 2017 sowie die Projekt-Dokumentation (OUDB 2021; OUDB 2017a; OUDB 2017b). Das annotierte Korpus liegt relational strukturiert als SQL-Datenbank vor, vgl. dazu das Datenbankschema in Abbildung 5.7. Die Korpusauswertung in Kapitel 6 dieser Arbeit basiert dementsprechend auf dem Annotationsschema des annotierten OUDB-Korpus, dessen einzelne Annotationskategorien im Folgenden kurz erläutert werden (vgl. auch OUDB 2017b). Eine vollständige Auflistung der Kategorien und der dazugehörigen Tag-Sets sind im Annotationsverzeichnis zu finden.

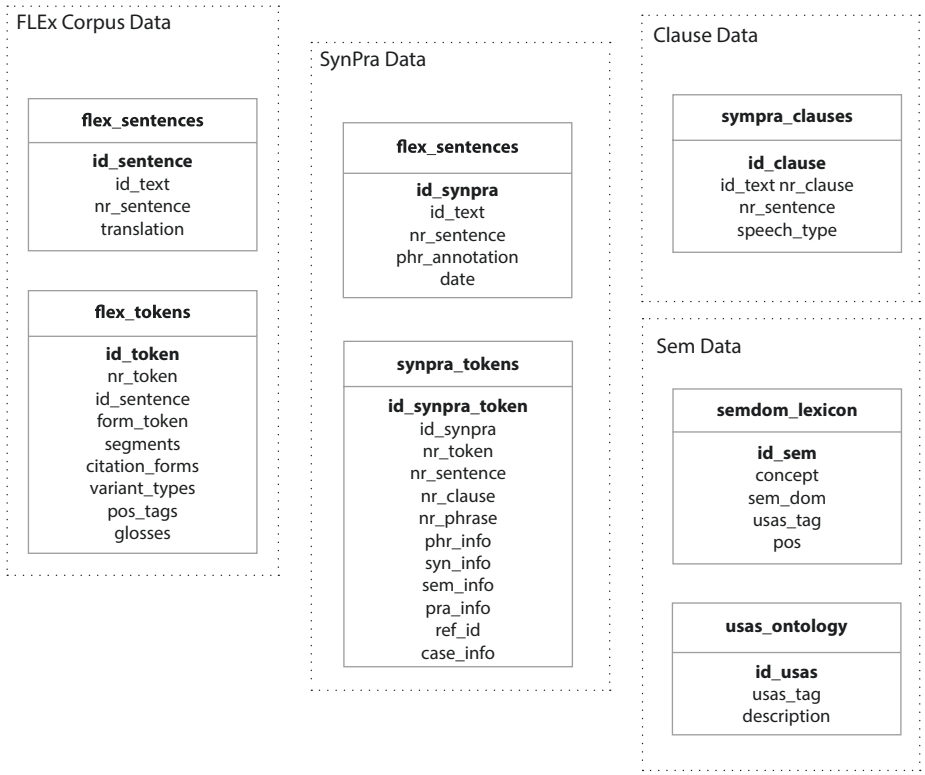


Abbildung 5.7: Relationales Datenbankschema des syntaktisch, pragmatisch und semantisch annotierten Korpus (angepasste und erweiterte Version von Wisiorek & Schön 2017: 388, Abb. 2)

5.3.1 Morphologische Annotationen

Die morphologische Auszeichnung der Korpusdaten erfolgte im Rahmen des OUDB-Projekts mit dem halbautomatischen Annotationstool FLEx (FieldWorks Language Explorer) nach einheitlichen Annotationskriterien (vgl. Wisiorek & Schön 2017). Die für die Analysen in dieser Arbeit relevanten morphologischen Auszeichnungen (u. a. für Subordinationsbackgrounding oder TAM-Foregrounding) sind **Wortarten** (s. Abschnitt „Part-of-Speech-Tags“ im Annotationsverzeichnis) und **morphologische Flexionskategorien** (s. Abschnitt „Morphologische Glossen“ im Annotationsverzeichnis).

5.3.2 Syntaktische Annotationen

Das Korpus liegt in einer flachen syntaktischen Annotation auf Phrasen-, Clause- und Satzebene vor, die im Rahmen des OUDB-Projekts über die Anwendung kaskadierender partieller Konstituentenregeln und einer Clause-Erkennungsheuristik gewon-

nen wurde; dieser Ansatz einer flachen, **partiellen Konstituentenanalyse** wurde aufgrund der intendierten, auf der syntaktischen Analyse basierenden Auszeichnung von informationsstrukturellen Größen gewählt – ähnlich der in der Informationsextraktion üblichen Anwendung eines partiellen Chunk-Parsings zur Identifizierung der zentralen Informationseinheiten in Texten (s. Manning & Schütze 1999: 376). Entsprechend sind auf der Phrasenebene zusätzliche Analyse-Elemente für Nullanaphern und Possessivsuffixe eingefügt, sodass ein anschließendes vollständiges Referenten-Tagging durchgeführt werden konnte (s. Wisiorek & Schön 2017: 389f.; Janda & Wisiorek & Eckmann 2017: 123ff.). Die so identifizierten syntaktischen Einheiten des Korpus sind gemäß ihrer funktionalen, semantischen und pragmatischen Rolle getaggt, und zwar unter Verwendung sprachspezifischer Heuristiken und anschließender manueller Nachkorrektur (s. Janda & Wisiorek & Eckmann 2017: 123ff.).

Konstituentenstruktur. Die Identifizierung syntaktischer Einheiten ist zentral für die Feststellung der Kodierung von TWM-Informationseinheiten (vgl. 3.1.1): Nominalphrasen kodieren Referenten, Verbalphrasen kodieren Relationen, Clauses kodieren Ereignisvorstellungen und Sätze kodieren miteinander gekoppelte Ensembles von Ereignisvorstellungen (sog. *scenarios*, s. Schulze 2000b). Im Folgenden werden die für diese Arbeit relevanten, zentralen Eigenschaften der syntaktischen Annotation des OUDB-Korpus vorgestellt (s. Abschnitt „Syntaktische Annotationsparameter" im Annotationsverzeichnis).

Die höchste syntaktische Analyseebene ist die **Satzebene**. Ein Satz besteht aus 1 bis n Clauses, die gewöhnlich durch das Vorhandensein eines finiten Verbs gekennzeichnet sind; dieses wird funktional als `PRED` getaggt (s. u.; in Nominalclauses entsprechend das Subjekt). Die Clauses eines Satzes sind durch Koordinationsrelationen (im Obugrischen meist ohne verbindende Konjunktion) oder Subordinationsrelationen verbunden, wobei im Obugrischen subordinierte Clauses – also einfache Sätze mit finitem Verb, die in einen Matrixclause eingebettet sind – selten anzutreffen sind (vgl. 5.1.2).

Auch auf **Phrasenebene** liegt eine flache Struktur ohne verschachtelte Phrasen vor; Attribute sind entweder als Teil der NP oder ggf., wenn sie eine Referenz enthalten, als eigenständige, annotierbare Einheit ausgezeichnet.

Die Labels von **Nominalphrasen** enthalten funktional-grammatikalische Informationen (etwa `locNP` als NP mit Lokalkasus); Postpositionalphrasen sind als `postP` gelabelt. Wie oben bereits ausgeführt, sind zur Ermöglichung einer vollständigen referentiellen Analyse die syntaktischen Einheiten im OUDB-Korpus um **Nullmorpheme** als Analyseeinheiten ergänzt (`zero` als über Agreement identifizierte Subjekt- bzw. Objekt-Argumentstelle) sowie die **Possessivsuffixe** (`px`) als eigenständig zu analysierende Einheiten abgetrennt (s. Janda & Wisiorek & Eckmann 2017; vgl. auch Janda 2015).

Auch **Verbalphrasen** enthalten eine funktionale Subdifferenzierung in ihren Labels; bei finiten VPs `okVP` für die objektive Konjugation, `passVP` für Passivformen und `finVP` für die subjektive Konjugation (Default-Fall des monopersonalen Agreements; s. 5.1.2). Da im Obugrischen Subordination primär über nicht-finite Verbformen realisiert wird, sind subordinierte Konstituenten entsprechend (und auch aufgrund der flachen syntaktischen Analyse) nicht als eingebettete Clauses analysiert, sondern explizit als subordinierte Konstituenten ausgezeichnet (`compC` als subordiniertes Komplement; `ptcpVP` als subordinierte attributive Einheit; `subC` als subordinierte adverbiale Einheit). Funktional werden diese subordinierten Einheiten als `SUBPRED` ausgezeichnet.[29]

Syntaktische Funktionen. Wie oben bereits besprochen, wird im OUDB-Korpus bei der Auszeichnung im verbalen Bereich differenziert zwischen dem Prädikat eines Clauses (`PRED`; meist ein finites Verb) und einem subordinierten Prädikat (`SUBPRED`). Im nominalen Bereich liegt als Annotation der syntaktischen Funktion eine Auszeichnung der **grammatischen Relationen** der referentiellen Einheiten vor (Subjekt `S`, Objekt `O`, Indirektes Objekt `IO` und Adverbial `ADV`). In dieser Arbeit erfolgt zusätzlich eine Subdifferenzierung im Subjektbereich, um die zu erprobende TWM-Operationalisierung so allgemein zu halten, dass sie auch für Sprachen anwendbar ist, die kein Akkusativ-Alignment aufzeigen: Es wird unterschieden zwischen dem Subjekt des intransitiven Satzes (S = Subjective) und dem agensartigen Argument (A = Agentive) des transitiven Satzes (vgl. Schulze & Sallaberger 2007), das in Akkusativsprachen (wie den obugrischen Sprachen) im transitiven Satz als dessen Subjekt fungiert, d. h. also als das in der Hierarchie grammatischer Relationen höchststehende Argument (s. Abschnitt „Funktionale Annotationsparameter“ im Annotationsverzeichnis).

5.3.3 Semantische Annotationen

Semantische Rollen. Mit Agens (`AG`) und Patiens (`PAT`) sind im Korpus semantische **Makrorollen** ausgezeichnet; diese Rollen sind in dieser Arbeit u. a. auch für die Differenzierung zwischen dem Subjekt des transitiven und dem des intransitiven Satzes relevant (Agentive vs. Subjective, vgl. 5.3.2). Im Recipient- und Lokativbereich werden verschiedene semantische Rollen differenziert (s. Tabelle 5.10 und Abschnitt „Semantische Rollen“ im Annotationsverzeichnis).

29 Infinitiv- und Partizip-Formen können auch als `PRED` auftreten, dann als `infVP` bzw. `ptcpVP` mit `PRED`-Funktionalauszeichnung.

Makrorolle	Tags
Agens	AG
Patiens	PAT
Recipient	REC, ADR, COM
Lokativ	LOC, GOAL, SOURCE, PATH, INST, TIME, MANNER, CAUSE, CONS, QUAL, DEG

Tabelle 5.10: Im Korpus getaggte semantische Rollen

Semantische Klassen. Die nominalen und verbalen Einheiten des Korpus sind – basierend auf ihren englischen Übersetzungen – mit dem UCREL Semantic Analysis System (USAS; s. Archer & Wilson & Rayson 2002) im Rahmen des OUDB-Projekts semantisch getaggt worden. In der Auswertung in Kapitel 6 werden diese USAS-Tags (s. Tabelle 5.11) aus der Datenbank ausgelesen und Subbereiche der USAS-Tags über in SQL-Abfragen implementierte Regeln (s. Auflistung 5.1) sowie manuelle Verblisten auf zentrale nominale **semantische Domänen** (Animate, Body Part, Human, Inanimate) und **Verbklassen** (Action&Process, Motion, Perception, Speech, State) abgebildet (vgl. Schulze 2019: 27f.; s. Abschnitt „Semantische Klassen" im Annotationsverzeichnis).

```
IF((usas_tag LIKE 'S2' OR usas_tag LIKE 'S3' OR usas_tag LIKE 'S4' OR usas_tag
    LIKE 'S7'),'HUM',
IF(usas_tag LIKE 'L3','INANIM',
IF(usas_tag LIKE 'L2','ANIM',
IF(usas_tag LIKE 'B1','BODY',
'INANIM'))))

IF(usas_tag LIKE 'Q_','SPEECH',
IF(usas_tag LIKE 'M8','STATE',
IF(usas_tag LIKE 'M_','MOTION',
IF((usas_tag LIKE 'X3' OR usas_tag LIKE 'X2'),'PERCEPT',
'ACT'))))
```

Auflistung 5.1: SQL-Regeln für die Abbildung von USAS-Tags auf semantische Klassen

A	B	C	E
General and abstract terms	the body and the individual	arts and crafts	emotion
F	**G**	**H**	**I**
food and farming	government and public	architecture, housing and the home	money and commerce in industry
K	**L**	**M**	**N**
entertainment, sports and games	life and living things	movement, location, travel and transport	numbers and measurement
O	**P**	**Q**	**S**
substances, materials, objects and equipment	education	language and communication	social actions, states and processes
T	**W**	**X**	**Y**
time	world and environment	psychological actions, states and processes	science and technology
Z			
names and grammar			

Tabelle 5.11: Semantische Hauptklassen des USAS-Tagsets (nach Archer & Wilson & Rayson 2002: 2); s. auch http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf für eine vollständige Liste mit Unterkategorien.

5.3.4 Referenzsemantische Annotationen

Referenten sind im syntaktisch annotierten OUDB-Korpus im Rahmen einer halbautomatischen Referenten-Annotation über **Referentenindizes** ausgezeichnet (für Details s. Janda & Wisiorek & Eckmann 2017: 123ff.; vgl. auch OUDB 2017b, insbesondere bzgl. der Auszeichnung von Gruppen). Die Nummerierung der Referenten folgt dabei primär der Reihung ihrer Ersterwähnung im Text. Die in der syntaktischen Annotation als Analyseeinheiten für Nullanaphern eingeführten `zero`-Elemente ermöglichen die Auszeichnung struktureller Bezugnahme, also nicht-overt kodierter Referentenerwähnungen.

5.3.5 Pragmatische Annotationen

Pragmatische Rollen. Die im Korpus vorliegende Auszeichnung der pragmatischen Funktion bezieht sich primär auf Einheiten des nominal-referentiellen Bereichs. Neben **Topik** (`TOP`) als Satzgegenstand für vorerwähnte, aktivierte Referenten (im Obugrischen primär über Nullmarkierung angezeigt) und **Fokus** (`FOC`) für neu eingeführte Referenten sind hier als markierte Parameter für die Aufmerksamkeitssteuerung insbesondere die folgenden, im OUDB-Korpus ausgezeichneten pragmatischen Funktionen relevant: `CTR` als **kontrastiver Fokus** (im Obugrischen pronominal markiert, also z. B. über Verwendung von Personalpronomina statt Nullmarkierung), `REPEAT` als unmittelbar nominal wiederholter Referent sowie `FRAME` als **Frame-Setting** (also die Einführung wichtiger Referenten als neue Diskurstopiks zu Beginn des Textes bzw. auch zu Beginn einzelner Textabschnitte); für Details bzgl. der

pragmatischen Annotation s. die Dokumentationen (OUDB 2021; OUDB 2017a; OUDB 2017b) sowie Janda & Wisiorek & Eckmann 2017: 124 (s. auch Abschnitt „Pragmatische Annotationsparameter“ im Annotationsverzeichnis).

Auszeichnung direkter Rede. Der textinterne diskursive Status von Clauses als ihr Vorkommen in Dialogphasen direkter Rede, der hier für die Berechnung des entsprechenden informationsstrukturellen Parameters relevant ist (textinterne Diskursstruktur), wurde im Korpus manuell getaggt; um Verwechslung mit der semantischen Klasse von Verben des Sprechens (`SPEECH`) zu vermeiden, wird das Feature des textinternen diskursiven Status in dieser Arbeit als `INFOSPEECH` getaggt.

5.3.6 Annotations- und Sprachbeispiele

Die besprochenen Annotationsebenen und -parameter können anhand von Interlinearversionen exemplifiziert werden: Jeder Absatz eines solchen satzbezogenen Sprachbeispiels umfasst dabei in der hier gewählten Darstellung einen Clause. In den sieben Zeilen der Interlinearversion sind folgende Ebenen ausgezeichnet:[30]

1. morphologisch segmentierte Worttokens, inkl. `zero` und `px`-Einheiten
2. Glossen
3. Phrasenstruktur
4. syntaktische Funktionen
5. semantische Funktionen
6. pragmatische Funktionen
7. Referentenindizes

30 Die semantischen Klassen der Einheiten erscheinen in dieser über die Export-Funktion der OUDB-Website generierten Ansicht nicht; diese Informationen sind aber als Teil des OUDB-Lexikons in der Datenbank abgespeichert (vgl. Abbildung 5.7).

Beispiel A. Das erste Sprachbeispiel aus dem khantischen `LittleBird`-Märchen 1314 exemplifiziert die verschiedenen Annotationsebenen und -parameter am Beispiel eines Satzes mit eingebetteter Konstruktion (diese nicht-finiten, meist partizipialen, subordinierten Konstruktionen des Obugrischen werden nicht als eigenständiger Clause analysiert, sondern als in den finiten Clause eingebettete nicht-satzwertige Konstruktionen; s. 5.1.2, 5.3.2):

(1) Text 1314, Satz 86 (PA-Korpus)

Ø	toβənə	βɑɬ-t-in		sɐːət	-Px
ø	so	live-PTCP.PRS-3DU		while	-PX
[zero]	**[subC**				**]** **[px]**
S	SUBPRED				AGR
PAT	TIME				AG
TOP	THL				TOP
1+2	–				1+2

jəppəɣ-nə	ojəɣt-iɣən
owl-LOC	find+[PST]-PASS.3DU
[locNP]	**[passVP]**
ADV	PRED
AG	–
FOC	FOC
25	–

While living like this, they were found by an owl.

Beispiel B. Das nachfolgende Sprachbeispiel des kurzen nordmansischen `Creation`-Textes 741, der aus nur vier Clauses besteht, zeigt, wie sich die in 5.1.2 erörterten, spezifischen sprachlichen Strategien der obugrischen Sprachen zur Kodierung von Topik-Kontinuität über Wortstellung und differentielles Agreement in den Annotationsdaten ausdrücken:

(2) Text 741, Satz 1 (NM-Korpus)

ta:xt	sʲa:rsʲ	patta-nl	ma:	xulʲiɣt-s	#
black-throated_loon	sea	bottom-ABL	earth	lift-PST[3SG]	
[NP]	**[locNP**	**]**	**[NP]**	**[finVP]**	
S	ADV		O	PRED	
AG	SOURCE		PAT	–	
FRAME	FRAME		FRAME	FRAME	
1	2		3	–	

Ø	jol_sʲalt-əs	#
ø	dive_down-PST[3SG]	
[zero]	**[finVP]**	
S	PRED	
AG	–	
TOP	FOC	
1	–	

Ø	su:p	ki:wərn	ma:	wi-s	#
ø	mouth	into	earth	take-PST[3SG]	
[zero]	**[postP**	**]**	**[NP]**	**[finVP]**	
S	ADV		O	PRED	
AG	GOAL		PAT	–	
TOP	FOC		REPEAT	FOC	
1	4		3	–	

Ø	Ø	nox_tot-as-te
ø	ø	bring_up-PST-SG<3SG
[zero]	**[zero]**	**[okVP]**
S	O	PRED
AG	PAT	–
TOP	TOP	FOC
1	3	–

The black-throated loon lifted up earth from the bottom of the sea: He dove down, put earth into his mouth, and brought it up.

6 Ergebnisse

Kapitelzusammenfassung

Die in Kapitel 4 eingeführten Repräsentations- und Klassifikationsmethoden zur gebrauchsbasierten Untersuchung textstruktureller Mustertypen für eine Text-Weltmodell-Mustererkennung werden in diesem Kapitel im Sinne einer Pretest-Fallstudie auf das in Kapitel 5 vorgestellte, zeitlich, räumlich und textsortenspezifisch geschichtete Korpus der beiden obugrischen Sprachen Mansi und Khanty angewendet, indem die in Kapitel 3 besprochenen TWM-Parameter einer kognitiven Texttypologie berechnet (Feature-Construction) und auf diesen basierende textstrukturelle Feature-Sets bzw. Sequenzfolgen extrahiert werden (Feature-Extraction; s. Abbildung 6.1).

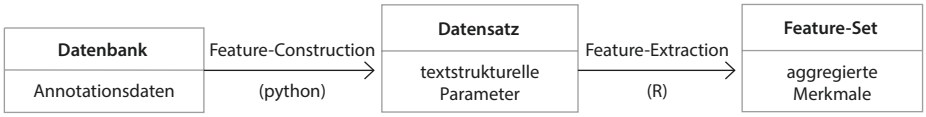


Abbildung 6.1: Vorgehen zu Erstellung von Feature-Sets

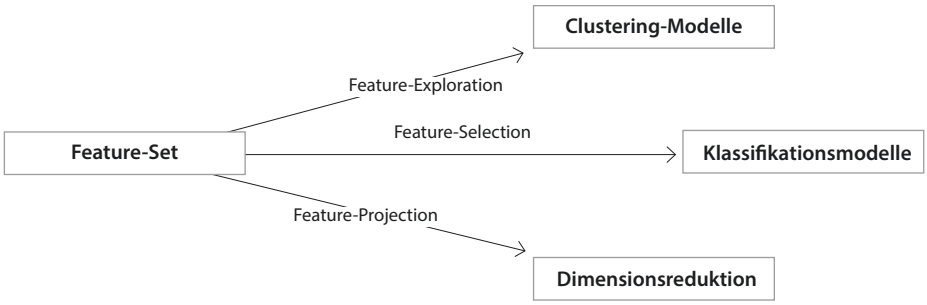


Abbildung 6.2: Übersicht der Auswertungsmethoden

Zur Rekonstruktion möglicher, den Texten des Korpus zugrundeliegender kognitiver Genre-Strukturmodelle werden anschließend auf diesen Daten Clustertypologien erstellt, indem durch agglomeratives hierarchisches Clustering Gruppen von Texten unterschiedlicher textstruktureller Mustertypen identifiziert werden (Feature-Exploration). Diese induktiven Gruppeneinteilungen können dann im Rahmen einer Clusterevaluation mit den in Kapitel 5 vorgestellten Apriori-Kategorisierungen abgeglichen werden. Des Weiteren wird ein Random-Forest-Klassifikator verwendet, um für die verschiedenen, in einem Feature-Set zusammengefassten TWM-Merkmale eine Gewichtung bzgl. ihrer Relevanz für die Gruppendifferenzierung sowohl für die Klassen dieser Apriori-Kategorisierungen als auch für die induktiv gefundenen

Clustertypologien zu erhalten (Feature-Selection). Außerdem wird für verschiedene Feature-Sets bzw. Sequenzfolgen durch geeignete Klassifikationsmethoden deren Fähigkeit zur Diskrimination der Apriori-Textklassen verglichen. Dabei kommen die in Kapitel 4 vorgestellten Visualisierungsmethoden zur Darstellung der Feature-Set-Analysen sowie der Sequenzanalysen zum Einsatz (Feature-Projection; s. Abbildung 6.2).

Bevor in den Abschnitten 6.2–6.5 die Ergebnisse der Analysen für die verschiedenen kognitiven Strukturbereiche vorgestellt werden, wird in Abschnitt 6.1 zunächst das konkrete, in dieser Arbeit durchgeführte Vorgehen der maschinellen Datenverarbeitung zum Aufbau textstruktureller Feature-Sets bzw. Sequenzrepräsentationen sowie deren weitere Verarbeitung in Clustering- und Klassifikationsverfahren unter Angabe der verwendeten Programmierumgebungen und -bibliotheken für die statistische Datenanalyse detailliert erläutert, um Sprachwissenschaftlern eine mögliche Replikation der Methodik zu erleichtern. Ebenso werden dort auch die Default-Werte für die Tuning-Parameter der verschiedenen Typen von Klassifizierungsmodellen und Visualisierungsmethoden angegeben, die entsprechend in den Auswertungen der folgenden Abschnitte Anwendung finden (sofern dort nicht explizit auf eine Verwendung abweichender Werte hingewiesen ist). Abschließend werden die Ergebnisse in den Abschnitten 6.6 und 6.7 einer Gesamtbewertung bzgl. der Eignung der untersuchten Parameter für eine TWM-Mustererkennung unterzogen und entsprechend diskutiert.

6.1 Angewandte Methoden textstruktureller Klassifizierung

6.1.1 Vorgehen zur Feature-Construction

Mit Methoden der **Feature-Construction** werden die quantitativen kognitiven Parameter (s. Kapitel 3) durch die Berechnung neuer Merkmale auf dem primären Datensatz von textstrukturellen Annotationsparametern des Korpus erzeugt; dazu gehört insbesondere die Generierung von Frequenzdaten, aber auch z. B. die Berechnung von Textabständen (etwa der referentiellen Distanz) durch konstruktive Operatoren. Für kategoriale Sequenzen ist im Allgemeinen keine Anwendung von Feature-Construction-Methoden notwendig, da diese direkt als *tag*-Folgen aus dem Primärdatensatz extrahiert werden können; ggf. ist aber eine numerische Rekodierung angebracht (Dummy-Encoding, vgl. 4.4.1.2). Für Partitur-Folgen werden Frequenzdaten einer Einheit bzgl. ihres Vorkommens in der Folge einer übergeordneten Einheit berechnet und anschließend extrahiert.

Zur Anwendung der Feature-Construction-Methoden werden zunächst die als Tokenlisten abgespeicherten Korpusdaten zusammen mit den primären Annotati-

onsdaten mit einem Python-Skript aus der SQL-Datenbank ausgelesen, in der die Daten des Korpus vorliegen (vgl. Abschnitt 5.3). Dieser Datensatz wird dann in eine adäquate Dictionary-Datenstruktur transformiert. Anschließend werden auf diesen Primärdaten für jede Strukturebene (Token, Phrase, Clause, Satz; Referenten, Ereignisvorstellungen) die oben diskutierten kognitiven Texttypologie-Parameter berechnet (z. B. die referentielle Distanz für jedes einzelne Vorkommen eines Referenten), indem **konstruktive Operatoren** (Addition etc., vgl. 4.1.2.1) auf die basalen grammatischen, semantischen und informationsstrukturellen Annotationskategorien angewendet werden. Pro Strukturebene wird als Ergebnis dieser Feature-Construction schließlich wieder ein tabular strukturierter Datensatz erzeugt, der um die für die jeweilige Untersuchungsgröße berechneten Parameter angereichert ist.

6.1.2 Vorgehen zur Feature-Extraction

Im Prozess der **Feature-Extraction** werden aus den in der Feature-Construction mit Python erzeugten Datensätzen kognitiver Textstruktur-Merkmale einer bestimmten linguistischen Strukturebene kompaktere Feature-Sets bzw. Sequenzfolgen extrahiert, die dann als textstrukturelle Repräsentationen Grundlage für die Auswertung mit Methoden der explorativen Datenanalyse und der automatischen Klassifizierung als mögliche TWM-Parameter sind. In dieser Arbeit geschieht diese analytische Verarbeitung der Daten mit dem Statistiktool R (R Core Team 2018).[1]

Für die Extraktion von Feature-Sets aus den Tokenlisten der Datensätze werden ggf. (insbesondere bei den globalen Textstruktur-Parametern) verschiedene **Filterungen** (Mappings, s. 4.1.2; vgl. Motoda & Liu 2002: 68f.) und **Aggregationen** durchgeführt, um aus den zuvor konstruierten Parametern die Feature-Werte zu berechnen.[2] Meist ist auch eine Textlängennormierung sowie z. T. auch eine Skalierung durchzuführen (vgl. Abschnitt 4.1). Ergebnis ist ein normiertes und ggf. skaliertes **Feature-Set** mit n Merkmalen (Spalten), das in einem Data-Frame abgespeichert wird; diese tabulare Datenstruktur enthält pro Zeile (d. h. pro repräsentiertem Textdokument) einen Merkmalsvektor, dessen Länge der Anzahl der Feature-Dimensionen (Spalten) ent-

1 Bei der Verwendung von Methoden, die für die Datenanalyse der Arbeit von zentraler Bedeutung sind, wird der entsprechende R-Code angegeben. Für Methoden, die nicht Teil des R-Kerns sind, wird deren Programmbibliothek in Klammern mitangeführt.

2 Dabei hat jede Strukturebene (und damit jeder entsprechende Datensatz) eine unterschiedliche Tokendefinition; so sind etwa Clauses die Einheit des entsprechenden Primärdatensatzes (Clause-Datensatz, s. Abschnitt 6.2).

spricht und der den Text als einen Datenpunkt im n-dimensionalen Merkmalsraum des Feature-Sets repräsentiert.[3]

Neben solchen Feature-Set-basierten Textstruktur-Repräsentationen werden auch **sequenzbasierte Repräsentationen** extrahiert (s. Abschnitt 4.4), welche die Textstruktur als lineare *tag*-Folge modellieren. Unterschiede in der Methodik für diese beiden Repräsentationsarten treten sowohl bzgl. der Extraktion der Daten auf (Sequenzextraktion, s. 6.1.2.4ff.), als auch bzgl. ihrer Weiterverarbeitung und Auswertung (Sequenzclustering und -klassifikation, s. 6.1.3.2ff.; 6.1.4.2ff.).

Im Einzelnen werden in dieser Arbeit konkret vier Typen textstruktureller Repräsentationen unterschieden, für die jeweils unterschiedliche Feature-Extraction-Methoden anzuwenden sind:

1. **Feature-Sets von globalen Textstruktur-Features**
2. **Textstrukturelle *bag*-Feature-Sets** (Bag-of-Tags, z. T. regional differenziert)
3. **Sequenzen** (*tag*-Folgen oder numerische Folgen, insb. **Partitur-Folgen**)
4. **Sequentielle *bag*-Feature-Set-Modelle** (Bag-of-Frequent-Tag-Patterns)

6.1.2.1 Extraktion, Normierung und Skalierung von globalen Feature-Sets

Die globalen Feature-Sets bestehen aus einer Menge an verschiedenen selbstständigen Merkmalen auf Textebene – die Berechnung bzw. Aggregation der Merkmale geschieht also pro Text; es handelt sich demnach um textbezogene Parameter mit jeweils merkmalsspezifischer Skala. Die **Textlängennormierung** dieser Merkmale (s. 4.1.3) geschieht u. a. über die Berechnung von Durchschnittswerten und Verhältniswerten bzw. über weitere Normierungen wie den Guiraud-Index (s. Unterabschnitt 3.2.2).

Die Werte der einzelnen Features werden bei ihrer Berechnung in R als Listen der Länge des Korpus abgespeichert (ein Wert pro Textdokument); anschließend werden diese zur Erstellung des Feature-Sets in einem Data-Frame als Datenmatrix kombiniert. Report 6.1.1 zeigt ein Beispiel für ein solches globales Feature-Set: Eine Zeile dieser Datenstruktur ist die Merkmalsvektor-Repräsentation eines Textdokuments im durch das Feature-Set aufgespannten Feature-Space (vgl. 4.1.1), entspricht also einem Datenpunkt in diesem n-dimensionalen Merkmalsraum. Bei diesen globalen Feature-Sets, die disparate textbezogene Merkmale kombinieren und dementsprechend unterschiedliche Skalen besitzen, ist eine Skalierung notwendig, um die Vergleichbarkeit der verschiedenen Dimensionen im Merkmalsraum zu gewährleisten.

3 Die Auswahl der in einem Feature-Set zusammengefassten Merkmale erfolgt dabei zum einen gemäß *domain knowledge* (s. 4.1.2.2; vgl. Motoda & Liu 2002: 70; Tan & Steinbach & Kumar 2006: 57), zum anderen datengestützt (d. h. hier orientiert an den Ergebnissen der Feature-Selection-Methoden, s. dazu insbesondere 6.6.1).

Diese **Skalierung** geschieht hier mittels z-Standardisierung (s. 4.1.3) über die `scale`-Funktion von R.

Text-ID	CL_ELAB	CL_COMPLEX	SENT_COMPLEX	RED	LEX_DENS
728	0.33	−0.04	0.43	−0.58	0.06
730	−0.03	0.19	0.55	−0.34	0.83
732	−0.18	0.27	0.55	0.11	0.84
741	−1.23	−1.30	−1.44	−1.50	−1.77
742	−0.65	−0.65	−0.35	−0.56	1.06
750	−0.85	−0.65	−0.32	1.75	1.24

Report 6.1.1: Skaliertes Feature-Set (Globale Parameter)

Folgende Parameter werden als Merkmale globaler Feature-Sets extrahiert:

- **Clause-Elaboration**; **Komplexität**; **Redundanz**; **lexikalische** und **referentielle Dichte**
- **Referentielle Explizitheit** und **Inferenz**; **nominale Elaboration**
- **Relationale Explizitheit** und **Inferenz**; **verbale Elaboration**

6.1.2.2 Extraktion und Normierung von Bag-of-Tags

Während die globalen Modelle ein Textdokument über eine Menge an disparaten, textweiten Struktureigenschaften wie das Type-Token-Verhältnis oder die lexikalische Dichte repräsentieren und sich damit die in entsprechenden Feature-Sets kombinierten, globalen Merkmale auf quantitative Eigenschaften des *Textes* beziehen, repräsentieren im Gegensatz dazu die Merkmale von **Bag-of-Tags**-Feature-Sets (s. 3.1.1; 4.1.1) quantitative Eigenschaften einer *Texteinheit* (z. B. die Textfrequenzen semantischer Klassen von Verben).[4]

Allgemein entsprechen also die Dimensionen des *bag*-Feature-Sets dem Vokabular der relevanten Eigenschaft einer Texteinheit (d. h. dem Vokabular der *tags* der entsprechenden Annotationskategorie, s. 3.1.1; 4.1.1).[5] Die Werte entsprechen im einfachsten Fall der Frequenz dieser Terme, also der Häufigkeit der verschiedenen Ausprägungen des relevanten Merkmals (vgl. Manning & Raghavan & Schütze 2009: 117). Als einfache **Textlängennormierung** kann hier etwa die relative Textfrequenz dienen (s. 4.1.3.1). Da in solchen *bag*-Modellen die Features als die verschiedenen

4 Beispielsweise bilden im Text *ich kam, ich sah, ich ging* die beiden vertretenen semantischen Klassen `MOTION` und `PERCEPTION` die Feature-Dimensionen einer *bag*-Textrepräsentation bzgl. der relationalen Einheiten des Textes, die sich als Abbildung auf die Textfrequenz darstellen lässt: `(MOTION:2, PERCEPTION:1)`.

5 Vgl. Aggarwal 2018: 2 für Bag-of-Words-Modelle: „In this case, the ordering of the words is not used in the mining process. The set of words in a document is converted into a *sparse multidimensional representation*, which is leveraged for mining purposes. Therefore, the universe of words (or *terms*) corresponds to the dimensions (or *features*) in this representation."

Ausprägungen der Eigenschaft einer Texteinheit eine gemeinsame Skala besitzen, ist eine **Skalierung** nur bei Kombination von verschiedenen *bag*-Modellen zu einem kombinierten Feature-Set notwendig (etwa beim kombinierten Distanz-Persistenz-Modell in 6.3.3).

Mitunter müssen bei diesen *bag*-Modellen fehlende Werte durch Feature-spezifische Defaultwerte ersetzt werden.[6]

Zusammenfassend werden im Aufbau solcher *bag*-Feature-Sets für kognitiv-textstrukturelle Einheiten eines bestimmten Typs (Referenten, Ereignisse usw.) textweite Frequenz- oder Durchschnittswerte von Merkmalen extrahiert, die für diese Einheiten im Feature-Construction-Prozess berechnet werden (z. B. die Distanz zwischen Referenten), wobei das Vokabular der Ausprägungen dieser syntaktischen, semantischen oder pragmatischen Eigenschaften (*tags*) die Merkmale eines solchen Bag-of-Tags-Feature-Sets bildet. Ggf. kann hier im Rahmen der Feature-Extraction auch noch eine Aggregation vorgenommen werden (etwa bzgl. der syntaktischen Funktion von Referenten bei der referentiellen Distanz, s. 6.3.1). Wie in 4.1.1 dargelegt, ist ein solches Bag-of-Tags-Modell eine rein textfrequenzbezogene Repräsentation bestimmter quantitativer Eigenschaften textstruktureller Einheiten, d. h. die Information zu deren linearer Anordnung im Text wird entfernt.

Folgende Parameter werden als Bag-of-Tags-Feature-Set modelliert:

- **Referentielle Distanz** und **Topik-Persistenz**
- **Topikalitätsquotient**
- **Ereignistypik**
- **Temporal-Sequencing-Stärke**
- **Pragmatische Typik**

6.1.2.3 Extraktion und Normierung von regional differenzierten Bag-of-Tags

Neben diesen textbezogenen Bag-of-Tags-Modellen (mit textweiter Aggregation pro Feature-Einheit) werden auch regional differenzierte Bag-of-Tags-Modelle untersucht, die durch eine **Partitionierung** des Textes in n Regionen die Distribution von textstrukturellen Parametern im Textdokument berücksichtigen und diese regionalen Informationen in die Merkmale integrieren. Dazu wird bei der Berechnung der aggregierten Werte pro textstruktureller Feature-Größe nicht pro Text aggregiert, sondern der Text wird in n Teile partitioniert und für jede dieser Textregionen

6 Vor Clustering und Klassifikation müssen in einigen Fällen fehlende Werte abgeschätzt werden. In dieser Fallstudie ist dies nur bei den Daten zur referentiellen Distanz bzw. Persistenz relevant (d. h. bei fehlenden Werten, wenn eine grammatische Relation nicht im Text vertreten ist) sowie beim Ereignistypik- und Pragmatiktypik-Feature-Set. Als Abschätzung werden die jeweiligen Defaultwerte verwendet (z. B. 0 bei Ereignistypik, da dieser Typ im Text nicht vorkommt). Hier handelt es sich eigentlich noch um einen Teil der Feature-Extraction bzw. -Construction, nicht um empirisch fehlende (nicht erhobene) Werte, deswegen ist die Anwendung von Imputationsmethoden zur Abschätzung fehlender Werte nicht notwendig.

werden die entsprechenden Werte berechnet. Die Regionen-Information wird zu einem Teil des Feature-Labels, es ergeben sich also Feature-Sets mit Länge/Umfang n x Anzahl Feature-Größen. Eine (hier regionenbezogene) Längennormalisierung geschieht über die Regionen-Frequenz bzw. den Regionen-Durchschnitt. Regional differenzierte Bag-of-Tags-Modelle sind also ein einfaches Modell zur Repräsentation sequentieller Information über eine Aufsplittung der von der linearen Anordnung abstrahierenden *bag*-Features nach Textregionen (die im Folgenden vorgestellten Sequenzmodelle sowie Bag-of-Frequent-Tag-Patterns-Modelle ermöglichen eine differenziertere Modellierung sequentieller Information).

Folgende Parameter werden als regional differenziertes Bag-of-Tags-Feature-Set modelliert:

- **Topik-Einführung**
- **Komplexitätsverlauf**

6.1.2.4 Extraktion von kategorialen Sequenzen

Neben Feature-Set-basierten Textstruktur-Repräsentationen werden in dieser Arbeit auch textstrukturelle Sequenzmodelle untersucht. Zur Extraktion textstruktureller kategorialer Sequenzmodelle als lineare *tag*-Folgen textstruktureller Werte (s. 4.4.1) werden pro Text aus den in der Feature-Construction erzeugten Datensätzen kategoriale *tag*-Folgen (**textstrukturelle Sequenzen**) als einfache Listen in R extrahiert und in einer Gesamtliste gespeichert (Datensatz von Sequenzen).

Dieser Datensatz wird für die Berechnung einer Distanzmatrix mit der `seqdef`-Funktion des `TraMineR`-Pakets in ein entsprechendes Sequenzobjekt umgewandelt, das (konzeptuell) eine Liste von nach den Text-IDs benannten kategorialen *tag*-Folgen enthält.

Statt einer Normierung bereits im Rahmen des Feature-Extraction-Prozesses wird bei kategorialen Sequenzen die Normierung erst im Analyse-Prozess bei der Berechnung der Distanzen zwischen den Sequenzen vorgenommen (anstelle einer Normierung von konstruierten numerischen Merkmalen werden hier die Distanzen zwischen den Sequenzen normiert).

```
R> seq.data.def <- seqdef(seq.data.matrix)
```

Auflistung 6.1: Erstellung eines Sequenzobjekts mit TraMineR

Folgende Parameter werden als kategoriale Sequenz modelliert:

- **Switch-Reference-Struktur**
- **Ereignisabfolgen**
- **Textinterne Diskursstruktur**

6.1.2.5 Extraktion und DTW-Normierung von numerischen und Partitur-Folgen

Eine Extraktion numerischer Sequenzen geschieht in dieser Arbeit sowohl bei der numerischen Rekodierung kategorialer Folgen mit Dummy-Variablen (s. 6.5.7; vgl. James u. a. 2017: 129f.) als auch im Rahmen der Extraktion von **Partitur-Folgen** (vgl. 3.1.2; 4.4.1.2), wo die Werte nach der Berechnung textfrequenzbezogener Merkmale von Sequenzelementen durch eine vorgelagerte Feature-Construction in einer Liste numerischer Folgen, benannt nach Text-IDs, abgespeichert werden.

Anders als bei Feature-Sets ist eine explizite Längennormierung solcher numerischen Folgen nicht notwendig, da diese implizit in der Berechnung der Dynamic-Time-Warping-Distanzmatrix im Rahmen der Klassifizierung erfolgt (s. 6.1.3.3). Im Rahmen der explorativen Datenanalyse solcher numerischer Sequenzen können die extrahierten Folgen ggf. auch geglättet werden (vgl. 4.4.1.2), hier mit der `smooth`-Funktion von R.

Folgender Parameter wird als DTW-normierte Sequenz modelliert: **Textinterne Diskursstruktur**.

6.1.2.6 Extraktion von Frequent-Tag-Patterns

Die Extraktion von sequentiellen *bag*-Modellen im Sinne von Bag-of-**Frequent-Tag-Patterns** verbindet Methoden der Sequenzextraktion mit anschließender Feature-Construction und -Extraction, indem aus textuellen Sequenzdaten extrahierte, wiederholt in Texten auftretende **lokale sequentielle Muster** als Merkmale in einem bag-Feature-Set zusammengefasst werden, also in einer Feature-basierten Repräsentation von Textstruktur über die Frequenzdaten dieser lokalen Muster in den Texten („bag of frequent subsequences", Aggarwal 2015: 503; vgl. auch 4.4.4).

Für die Extraktion von Frequent-Patterns bietet das `TraMineR`-Paket verschiedene Funktionen an (s. Gabadinho u. a. 2011; Studer & Ritschard 2016; Ritschard & Bürgin & Studer 2013):

1. **Extraktion von Sequenzen bzw. Übergängen**:
   - `seqdef {TraMineR}`
   - `seqecreate {TraMineR}`
2. **Frequent-Pattern-Feature-Construction**: `seqefsub {TraMineR}`
3. **Frequent-Pattern-Feature-Extraction**: `seqeapplysub {TraMineR}`[7]

Zunächst erfolgt eine Sequenzextraktion mit der `seqdef`-Funktion des `TraMineR`-Pakets; anschließend werden aus diesen Zustandsfolgen (z. B. Ereignistyp-Folgen der Art `-MOTION-MOTION-ACT-MOTION-`)[8] mit der `seqecreate`-Funktion des

7 Dokumentationen unter http://traminer.unige.ch/doc/seqecreate.html, http://traminer.unige.ch/doc/seqefsub.html und http://traminer.unige.ch/doc/seqeapplysub.html (abgerufen am 03.09.2022).
8 Das ist die Ereignistyp-Sequenzfolge von Text 741; vgl. Interlinearversion 2 in 5.3.6.

TraMineR-Pakets **Übergangsfolgen** extrahiert, also Sequenzen der Übergänge von einem Zustand in einen anderen, im Beispiel: (MOTION>ACT)-(ACT>MOTION).[9]

Durch die Betrachtung von solchen Zustandsänderungen (auch: *transitions* oder *transition events*)[10] erreicht man eine Abstraktion von der Verweildauer in einem bestimmten Zustand.[11] Als Grundlage für die Extraktion von häufigen Teilsequenzen reduziert die Verwendung solcher Übergangssequenzen die Länge der Sequenzen (und damit deren Komplexität) durch die Integration von sequentiellen Abfolgeinformationen in die Sequenz-Grundeinheiten, also durch Verwendung komplexer Labels wie (MOTION>ACT).

Für die im Folgenden mit diesem Modell operationalisierte Ereignisabfolge werden also nicht einfache *tag*-Sequenzen (Verbklassen-*tags* bei der Ereignisabfolge) als Ausgangspunkt genommen, sondern Übergangssequenzen erstellt und auf diesen eine Extraktion der häufigsten Verbklassen-Übergänge durchgeführt. Damit ist dies also genauer ein Bag-of-Frequent-Tag-Transitions-Modell. Der Grund für die Verwendung von häufigen Übergangssequenzen als Merkmale ist, dass die so erreichte Modellierung Blöcke (Chunks) der Übergänge verschiedener Ereignistypen repräsentiert (vgl. 3.7.3; 4.4.4) und damit ein verallgemeinertes Modell darstellt, das von der Verweildauer in den einzelnen Zuständen absieht. Im Gegensatz zu einfachen sequentiellen Mustern im Sinne von Zustands-N-Gramm-Mustern (vgl. Legallois & Charnois & Larjavaara 2018a: 6f.) abstrahieren diese Übergangs-N-Gramme also von der Länge der Sequenzen eines Zustands und betrachten nur die Abfolge der Zustandsänderungen.

```
trans.list <- seqecreate(seq.data.def)

fsubseq <- seqefsub(trans.list, pmin.support=0.67, max.k=3,
     constraint=seqeconstraint(max.gap=1))

msubcount <- seqeapplysub(fsubseq, method="presence")
featset.freq.pattern <- as.data.frame(msubcount)
```

Auflistung 6.2: Extraktion von Frequent-Transition-Patterns mit TraMineR

Feature-Construction. Die Extraktion von häufigen Teilsequenzen von Übergangsfolgen (s. Gabadinho u. a. 2011: 21f.) als Features eines Bag-of-Frequent-Transitions-

9 Die Zustandsfolge des Textes 741 der Länge 4 wird hier also auf eine Übergangsfolge der Länge 2 reduziert (wenn man von Start- und Endzustand absieht).

10 Vgl. Gabadinho u. a. 2009: 16; Gabadinho u. a. 2011: 3; Ritschard & Bürgin & Studer 2013.

11 Das Objekt für Übergangssequenzen des TraMineR-Pakets speichert die Information zur Verweildauer mit ab (-1- usw.); diese Information wird hier in der Feature-Extraction häufiger Übergangsteilsequenzen von Ereignistypen aber nicht verwendet.

Modells geschieht mit der `seqefsub`-Funktion des `TraMineR`-Pakets.[12] Als zentraler Parameter bestimmt `min.support` die Mindestanzahl an Objektsequenzen, in denen eine Teilsequenz vorkommt (den **Support**; s. als Beispiel Report 6.1.2; `Count` ist hier die Anzahl von Texten mit Support). Alternativ kann mit `pmin.support` auch die entsprechende relative Häufigkeit an Objektsequenzen mit Vorkommen einer Teilsequenz angegeben werden. Der minimale Support wird für die häufigen Ereignistyp-Übergänge mit 0.67 sehr hoch angesetzt, um die für das Korpus typischen, d. h. in den meisten Texten vorkommenden Muster zu finden, die den Großteil (also zwei Drittel) der Texte des Pretest-Korpus obugrischer Volkserzählungen auszeichnen (und um damit auch die Texte herauszufiltern, die in dieser Hinsicht nicht typisch sind).

Der `max-gap`-Parameter von `seqeconstraint`[13] regelt die Beschränkung des Zeitabstands für die Subsequences (s. 4.4.4). Da hier nur zusammenhängende Muster untersucht werden sollen (also **Substrings**, vgl. Abschnitt 4.4), wird er auf 1 gesetzt, d. h. zwei Events dürfen maximal eine Zeiteinheit entfernt sein und dürfen somit nur direkt aufeinanderfolgen. Es gibt dann folglich keine ‚Lücken', wie es bei Teilsequenzen im Allgemeinen möglich ist (so etwa beim Einsatz als Methodik in sozialwissenschaftlichen Biographie-Analysen wie in Beispieldaten des `TraMineR`-Pakets, s. Gabadinho u. a. 2009; Gabadinho u. a. 2011). Stattdessen sollen hier häufige **Übergangs-N-Gramme** untersucht werden; diese werden durch `max.k = 3` beschränkt auf maximal Trigramme (bei dem angesetzten hohen `pmin.support` werden im Korpus aber N-Gramme höherer Ordnung auch nicht gefunden).

	Support	Count	subseq
1	0.88	30.00	(ACT>MOTION)
2	0.85	29.00	(MOTION>ACT)
3	0.74	25.00	(SPEECH>ACT)
4	0.71	24.00	(ACT>MOTION)-(MOTION>ACT)
5	0.68	23.00	(MOTION>ACT)-(ACT>MOTION)

Report 6.1.2: Ergebnis der Extraktion (Häufige Ereignisübergänge)

Text-ID	(ACT>MOTION)	(MOTION>ACT)	(SPEECH>ACT)
728	3.00	4.00	0.00
730	7.00	5.00	1.00
732	10.00	10.00	4.00
741	1.00	1.00	0.00
742	6.00	5.00	2.00
750	21.00	13.00	18.00

Report 6.1.3: Ausschnitt des Feature-Sets mit Frequenzzählung (Häufige Ereignisübergänge)

Feature-Extraction. Schließlich wird mit der `seqeapplysub`-Funktion des `TraMineR`-Pakets (s. auch Gabadinho u. a. 2009) das Vorkommen der extrahierten Parameter häufiger Übergangssequenzmuster in den durch diese häufigen Übergangsse-

12 Der in der `seqefsub`-Funktion implementierte Algorithmus ist eine Präfix-Baum-basierte Suche (s. Masseglia & Teisseire & Poncelet 2004).

13 Dokumentation unter http://traminer.unige.ch/doc/seqeconstraint.html (abgerufen am 03.09.2022).

quenzen repräsentierten Textobjekten bestimmt und so das Feature-Set mit diesen Frequent-Patterns als Features erstellt. Als Zählmethode kann entweder eine Frequenzzählung (`method=count`, vgl. Report 6.1.3) oder eine Vorkommenszählung (`method=presence`, vgl. Report 6.1.4) angewendet werden.[14]

Text-ID	(ACT>MOTION)	(MOTION>ACT)	(SPEECH>ACT)	(ACT>MOTION)-(MOTION>ACT)	(MOTION>ACT)-(ACT>MOTION)
728	1.00	1.00	0.00	1.00	1.00
730	1.00	1.00	1.00	1.00	1.00
732	1.00	1.00	1.00	1.00	1.00
741	1.00	1.00	0.00	0.00	1.00
742	1.00	1.00	1.00	1.00	0.00
750	1.00	1.00	1.00	1.00	1.00

Report 6.1.4: Feature-Set Presence/Absence (Häufige Ereignisübergänge)

Durch die Verwendung einer Presence-Absence-Zählung kann hier auf eine Längennormierung verzichtet werden, da diese Zählung eben das Vorhandensein bzw. Fehlen der häufigsten Muster als Feature relevant macht und die Stärke vernachlässigt.[15]

Alternativ kann das Kosinus-Distanzmaß mit Frequenzwerten (`method=count`) verwendet werden, da dieses von der Länge abstrahiert, indem nur die Richtung bewertet wird. Eine Testauswertung zeigt im Vergleich zu dem hier gewählten Ansatz mit euklidischem Distanzmaß und mit Presence-Absence-Kodierung sowie hohem `min.support` ein ähnliches Resultat im Clustering mit Kosinus-Distanzmaß und 10%-`min.support` (und `method=count`). Hinsichtlich des Einsatzes in dem Gesamt-Feature-Set wird aber die Presence-Absence-Kodierung vorgezogen, da hier normierte Features vorliegen.

Folgender Parameter wird in dieser Arbeit als ein solches Frequent-Pattern-Modell modelliert: **Häufige Ereignisübergänge** (also typische Ereignisabfolgemuster als häufige N-Gramme von Verbklassen-Übergängen).

14 In der Auswertung dieser Arbeit wurde zunächst die `count`-Zählung verwendet und die 0/1-Kodierung für Presence/Absence nachträglich durchgeführt.

15 Ereignistypmuster sind mesostrukturelle Parameter, also Einheiten oberhalb des Satzes, die Ereignisblöcke (Chunks) bilden; entsprechend wird durch die Operationalisierung u. a. über *transitions* von Einzelereignissen abstrahiert (d. h. von der Verweildauer in den Zuständen). Deshalb ist auch die Abstraktion von den Frequenzen des Auftretens gerechtfertigt, da z. B. in kurzen Texten nur wenige solcher Übergänge möglich sind.

6.1.3 Vorgehen zur Feature-Exploration

6.1.3.1 Hierarchisches Clustering textstruktureller Feature-Sets

Die explorative Datenanalyse von Feature-Sets durch agglomeratives hierarchisches Clustering (s. Abschnitt 4.2 und 6.1.3) ist ein zweistufiger Prozess:

1. Die Berechnung der **Distanzmatrix** als Menge aller paarweisen Abstände zwischen den Datenpunkten wird mit der `dist`-Funktion von R durchgeführt.
2. Die **agglomerative** Berechnung der Abstände zwischen Gruppen von Datenpunkten (Clustern) geschieht mit der `hclust`-Funktion von R.

Für die Berechnung der Distanzmatrix als Sammlung der paarweisen Abstände zwischen den Datenpunkten ist ein geeignetes Distanzmaß festzulegen (`method`-Parameter von `dist`); als Default wird in dieser Arbeit hier für das Clustering von Feature-Sets das **euklidische Distanzmaß** verwendet (s. 4.1.4). Ebenso ist für die Berechnung von Abständen zwischen Clustergruppen im Clustering eine geeignete Agglomerationsmethode festzulegen (`method`-Parameter von `hclust`); als Standard-Agglomerationsmethode für Feature-Sets wird **Complete-Linkage** gewählt (vgl. 4.2.2); bei kategorialen Sequenzanalysen wird standardmäßig die **Minimum-Variance**-Agglomerationsmethode von Ward verwendet (vgl. 4.2.2; s. auch 6.1.3.2).

Mit der `dist`-Methode wird jeweils für zwei Merkmalsvektoren gemäß des gewählten Distanzmaßes (vgl. 4.1.4) deren Abstand berechnet und dieses Resultat in einer Kreuzmatrix abgespeichert; neben den von der `dist`-Methode zur Verfügung gestellten Distanzmaßen werden weitere Distanzmaße wie die DTW-Distanz mit eigenen Distanzfunktionen bzw. unter Zuhilfenahme entsprechender R-Pakete definiert (`dtw`, Giorgino 2009).

```
featset <- data.frame(rbind(c(1,2,0),c(1,4,0),c(0,2,3),c(5,5,5)))
print(featset)
>  X1 X2 X3
> 1 1 2 0
> 2 1 4 0
> 3 0 2 3
> 4 5 5 5
```

Auflistung 6.3: Feature-Set als Datenmatrix

```
distmatrix <- dist(featset, method="euclidean")
print(distmatrix)
>       1      2      3
> 2 2.000000
> 3 3.162278 3.741657
> 4 7.071068 6.480741 6.164414
```

Auflistung 6.4: Berechnung Distanzmatrix: dist

Die berechnete Distanzmatrix dient der Clustering-Funktion `hclust` als Input. Als hierarchisch-agglomerativer Clustering-Algorithmus setzt diese *bottom-up*-Methode (s. Abschnitt 4.2) zunächst jeden Datenpunkt als eigenen Cluster; dann werden die beiden gemäß des Agglomerationsmaßes ähnlichsten Cluster zu einem neuen Cluster zusammengefasst (Agglomeration).[16]

```
cluster <- hclust(distmatrix, method="single")
print(cluster$height)
> 2.000000 3.162278 6.164414
print(cluster$order)
> 4 3 1 2
```

Auflistung 6.5: Agglomeratives Clustering: hclust

Anschließend kann das Ergebnis des hierarchischen Clusterings in eine bestimmte Anzahl von Datengruppen eingeteilt werden, also eine **Clustertypologie** erstellt werden (s. Manning & Raghavan & Schütze 2009: 379f.). Dazu wird das `hclust`-Objekt der `cutree`-Funktion übergeben und die gewünschte Anzahl an Gruppen spezifiziert; diese kann z. B. über die Cluster-Qualitätsmaße wie die Average-Silhouette-Width bestimmt werden (s. 4.2.3). Das Ergebnis eines agglomerativen Clusterings kann dann als Baumstruktur visualisiert werden (Dendrogramm, für ein Beispiel s. Plot A.1), wobei die Blätter die einzelnen Datenpunkte repräsentieren und jeder Knoten eine binäre Gruppierung darstellt.[17]

Zur Beurteilung eines Clustering-Resultats können Cluster mit automatischen Methoden gelabelt werden (s. Manning & Raghavan & Schütze 2009: 396). In der Auswertung des obugrischen Korpus ist aufgrund der hier gewählten Korpusgröße auch das Labeling jedes Textobjektes mit seinen Metadaten möglich; in diesen Cluster-Labels sind die folgenden Informationen enthalten (s. 5.2.2 und 5.2.4): Titel-Code, Text-ID, Dialektkürzel, Sammlung, Textlänge (in Sätzen), `COMM_SIT`-, `DISC_STRUCT`-, `GENRE`-Klassen. Mit der `cutree`-Funktion wird auf dieser Baumstruktur nun die Clustertypologie erstellt, indem Teilbäume gemäß der gewünschten Anzahl an Clustern so ausgewählt werden, dass deren Knoten sich im Dendrogramm unterhalb einer bestimmten Höhe h (= Abstand zwischen den Clusterelementen) befinden; alle Elemente in diesen so ausgewählten Clustern haben also zueinander einen geringeren Abstand als h. Resultat von `cutree` ist eine Liste mit Gruppennummern, die

16 Als Agglomerationsmethoden (*linkage*-Typen) stellt `hclust` die oben (Abschnitt 4.2.2) besprochenen Clusteringmaße `complete`, `single`, `average` und `ward` (in zwei Varianten, `ward.D` und `ward.D2`) zur Verfügung.

17 Mit der `plot`-Methode für `hclust`-Objekte kann das Clusterdendrogramm als Baumrepräsentation dargestellt werden. Der Abstand zwischen zwei Clustern (gemäß des Agglomerationsdistanzmaßes der Clusteringmethode) kann hier aus der Höhe, auf der sie vereinigt sind (Knoten), abgelesen werden, vgl. 4.2.

die Gruppenzugehörigkeit der Texte angeben (etwa die Clustergruppen 1 und 2 in Abbildung 6.3; vgl. auch Gruppennummern im Dendrogramm Plot A.1).[18]

```
groups <- cutree(cluster, k=2)
print(groups)
> 1 1 1 2

plot(cluster)
rect.hclust(cluster, k=2, cluster=groups)
```

Auflistung 6.6: Erstellung Clustertypologie: cutree, rect

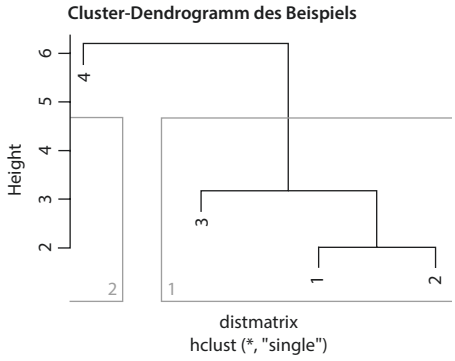


Abbildung 6.3: Dendrogramm des Clustering-Resultats

Die Auswahl der zu untersuchenden Clusteranzahl kann sich dabei einerseits empirisch an den Daten orientieren (Silhouette-Methode und Cluster-Qualitätsmaße zur Bestimmung der optimalen Clusteranzahl, vgl. 4.2.3), andererseits zum Vergleich mit einer Apriori-Kategorisierung (externe Evaluation) auch an der Anzahl der mit dem Ergebnis zu vergleichenden Klasseneinteilungen (vgl. Reiter & Frank & Hellwig 2014: 597). Als Standard wird hier entsprechend der zu untersuchenden Genre-Typisierung eine Clusteranzahl von drei verwendet, gemäß der Annahme von zwei narrativen Subgenres (Tier- vs. Zaubermärchen) und einer Gruppe peripherer Texte (ggf. wird diese Zahl beim Vorliegen von Ausreißern angepasst).

6.1.3.2 Clustering kategorialer Sequenzen

Für das Clustering von kategorialen Wertfolgen wird mit der `seqdist`-Funktion des `TraMineR`-Pakets eine Edit-Distance-Matrix mit paarweisen Abständen zwischen den Sequenzen auf den in einem `seqdef`-Zustandssequenz-Objekt gespeicherten Sequenzdaten berechnet (eine Liste von *tag*-Folgen; s. 4.4.2).

Wie in Kapitel 4 besprochen, können dann für diese Distanzmatrix kategorial-sequentieller Textstrukturrepräsentation die eben vorgestellten Clusteringmethoden auch für das Sequenzclustering Anwendung finden. Aufgrund des Formats der durch das `TraMineR`-Paket berechneten Distanzmatrix wird für das agglomerative Cluste-

18 Die Clustergruppen der Clustertypologie werden im Dendrogramm als Rechtecke eingezeichnet und ihre von `cutree` berechnete Gruppennummer angegeben (die Nummerierung folgt der Ordnung der Text-IDs, d. h. Cluster 1 enthält den Text mit der kleinsten ID usw.).

ring von kategorialen Sequenzdaten allerdings statt `hclust` die `agnes`-Funktion des `cluster`-Pakets verwendet.

Als Agglomerationsmethode für die **Sequenz-Clusteranalyse** wird hier analog zum Vorgehen bei Gabadinho u. a. (2011: 32; 2009: 101) **Wards** Methode minimaler Varianz verwendet (s. 4.2.2 und 4.4.2.1).[19] Als zugrundeliegendes sequentielles Distanzmaß wird Optimal-Matching (OM) verwendet; dieses Edit-Distance-Maß berücksichtigt zwei Transformationsoperationen (vgl. 4.4.2.1), nämlich *indel* (*insertion/deletion*) und *substitution* (Gabadinho u. a. 2011: 25f.; s. auch Studer & Ritschard 2016).

Für beide Operationen müssen die Kosten festgesetzt werden, also die Gewichtung, mit der die jeweilige Operation in den Gesamt-Editierabstand eingeht; dabei gilt: „Setting a high indel cost relatively to substitution costs favors substitutions while low values favor indels" (Gabadinho u. a. 2011: 26). Für *indel* wird hier ein niedriger konstanter Standardwert von 1 verwendet („Usually the indel cost is set as a constant independent of the concerned position and state", Gabadinho u. a. 2011: 26), d. h. es werden *indels* (Einfügungen oder Löschungen) bevorzugt, was wegen der Daten mit Sequenzen stark unterschiedlicher Längen sinnvoll ist, vgl. Gabadinho u. a. 2011: 28: „Favoring insertions and deletions reduces the importance of time shifts in the comparison, while favoring substitutions gives more importance to position-wise similarities."

Für die Substitutionskosten wird mit der `seqsubm`-Funktion des `TraMineR`-Pakets eine Substitution-Cost-Matrix mit der Methode `TRATE` berechnet.[20] Dabei werden die *transition rates* (`TRATE`) zwischen Zuständen des Sequenzalphabets berücksichtigt, d. h. die in den Daten gegebene Wahrscheinlichkeit, dass ein Zustand in einen anderen übergeht, dass also Zeichen s_i auf Zeichen s_j folgt: $p(s_i|s_j)$, und umgekehrt: $p(s_j|s_i)$ (s. Gabadinho u. a. 2011: 26). Diese beiden Werte werden zur Berechnung des `TRATE`-Kostenwertes von dem `CONSTANT`-Standard-Kostenwert 2 subtrahiert (s. Gabadinho u. a. 2011: 26); das heißt, je häufiger zwei Zustände in den Daten aufeinanderfolgen, desto geringer sind die Kosten für deren Ersetzung.

19 Die `hclust`-Implementierung der Ward-Clusterdistanz entspricht der ursprünglichen Version dieses Agglomerationsmaßes von Ward 1963; vgl. Murtagh & Legendre 2014; s. die Hinweise in R Core Team 2020.

20 Für Details s. Gabadinho u. a. 2011: 26 und die Dokumentation unter http://traminer.unige.ch/doc/seqcost.html (abgerufen am 31.10.2021).

```
submat <- seqsubm(data, method = "TRATE")
dist <- seqdist(data, method = "OM", indel = 1,sm = submat, norm = "maxlength")

cluster <- as.hclust(agnes(dist, diss = TRUE, method = "ward"))
```

Auflistung 6.7: Berechnung OM-Distanzmatrix und Clustering auf Sequenzen

Für die **Längennormalisierung** der Distanzen wird *maxlength* verwendet (s. 4.4.2.1); das ist auch das Standard-Normalisierungsverfahren der `seqdist`-Funktion für Optimal-Matching-Distanzen. Dabei geschieht die Normierung der Distanz zwischen zwei kategorialen Sequenzen über die Länge der längeren der beiden Sequenzen (s. Gabadinho u. a. 2011: 29: „Abbott's normalization, which consists of dividing the distance by the length of the longest of the two sequences"; vgl. Abbott & Tsay 2000).

6.1.3.3 Clustering von DTW-normierten Sequenzen

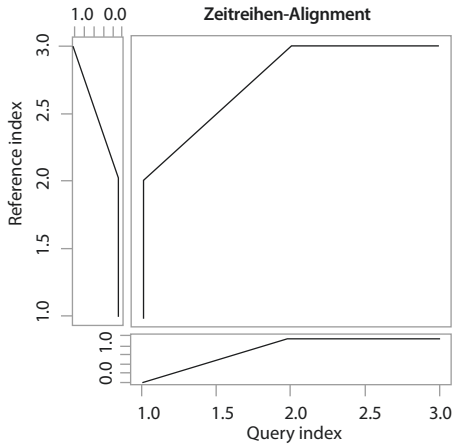


Abbildung 6.4: Alignment der Beispielzeitreihen mit Dynamic-Time-Warping

Zur Berechnung einer Distanzmatrix von paarweisen Dynamic-Time-Warping-Distanzen zwischen den numerischen Partitur-Folgen wird das `dtw`-Paket (Giorgino 2009) verwendet. Dazu wird entsprechend der `dtw`-Dokumentation[21] die Definition einer Distanzfunktion `dtwOmitNA` als Distanz-Methode für die `dist`-Funktion hinzugefügt, die den DTW-Algorithmus schrittweise auf jeweils zwei Sequenzen anwendet und die berechnete normalisierte **DTW-Distanz** wiedergibt. Anschließend wird diese Distanzmatrix der DTW-Abstände zwischen den durch numerische Folgen repräsentierten Textdokumenten an die `hclust`-Funktion für ein hierarchisches Clustering übergeben. Als Agglomerationsmethode wird beim DTW-Clustering analog zum Clustering von Feature-Sets **Complete-Linkage** verwendet sowie zum Vergleich mit dem OM-Clustering auf kategorialen Sequenzen Wards Agglomerationsmethode.

21 Dokumentation unter http://dtw.r-forge.r-project.org/ (abgerufen am 31.10.2021).

```
query <- c(0,1,1)
template <- c(0,0,1)
alignment <- dtw(query, template, keep=TRUE)
plot(alignment, type="threeway")
```

Auflistung 6.8: Paarweise Berechnung DTW-Distanz

```
cluster_data <- dataset.ref.switch
dist <- dist(cluster_data, method = "dtwOmitNA")
cluster <- hclust(dist, method="complete")
```

Auflistung 6.9: Clustering mit DTW-Distanz-Matrix

6.1.3.4 Clustering von Frequent-Patterns-Feature-Sets

Für das Clustering von Feature-Sets mit Teilsequenzen als Merkmalen können grundsätzlich dieselben Methoden wie für das Clustering der anderen Feature-Set-Typen eingesetzt werden. Allerdings kann hier aufgrund der automatischen Extraktion der N-Gramme durch das Sequential-Pattern-Mining (s. 4.4.4) ein potentiell sehr großes Feature-Set entstehen (abhängig von den Daten und der Operationalisierung), sodass hier ggf. folgende Punkte berücksichtigt werden müssen:

- die Qualität des Clusterings kann mit der Größe des Feature-Sets variieren, ist also abhängig von `(p)min.support` (s. 6.1.2.6)
- ggf. ist ein Clustering eines Subsets angebracht (mit Feature-Subset-Selection, s. 6.1.4)

```
cluster_data <- featset.ev.seq.norm.event
dist <- dist(cluster_data, method = "euclidean")
cluster <- hclust(dist, method="complete")
```

Auflistung 6.10: Clustering von Feature-Sets mit TraMineR-Teilsequenzen

6.1.3.5 Berechnung von Clusterevaluationsmaßen

Hopkins-Statistik. Der Wert der Hopkins-Statistik H als Maß zur Bestimmung der Cluster-Tendenz eines Datensatzes (s. 4.2.3) wird mit Hilfe der entsprechenden Funktion des `factoextra`-Pakets (Kassambara & Mundt 2017) bestimmt (`get_clust_tendency`); in der verwendeten Version (1.0.5) wird der Wert von H gemäß der Formel in 4.2.3 berechnet; d. h. je näher der Wert an 0, desto stärker ist die Cluster-Tendenz. Da für die Berechnung mit `get_clust_tendency` ein Feature-basierter Datensatz Voraussetzung ist, wird die Hopkins-Statistik nur für Feature-Sets berechnet, nicht für Sequenzdaten; in diesen Fällen werden andere Cluster-

Qualitätsmaße wie der Wert der Average-Silhouette-Width herangezogen, die mit dem R-Paket `WeightedCluster` (Studer 2013) berechnet werden können.

Optimale Clusteranzahl. Die Bestimmung der optimalen Clusteranzahl über die **Elbow**- bzw. die **Silhouette**-Methode (s. 4.2.3) erfolgt für Feature-Sets mittels der entsprechenden Funktion des `facto-extra`-Pakets (`fviz_nbclust`); für die Sequenz-Clusteranalysen wird die Average-Silhouette-Width mit dem `WeightedCluster`-Paket als eines der dort implementierten Cluster-Qualitätsmaße bestimmt.

Rand-Index. Als externes Maß zum Vergleich der Übereinstimmung von gefundenen Clustergruppen mit den Apriori-Klassen der Textsortenkategorisierung wird der **Adjusted-Rand-Index** (s. 4.2.3) mit dem R-Paket `pdfCluster` (Azzalini & Menardi 2014) berechnet.

6.1.4 Vorgehen zur Feature-Selection

6.1.4.1 Feature-basierte Klassifikation mit Random-Forest

Für die Klassifikation von Feature-Sets bzw. Sequenzen wird zunächst die Datenmatrix um eine Klassenlabel-Attributspalte ergänzt.

```
classdata <- featset
classdata$GENRE <- genrelist
print(head(classdata))
    ACT MOTION PERCEPT SPEECH GENRE
728  0.37 0.40 0.11 0.00 priv
730 0.33 0.32 0.03 0.03 priv
732 0.37 0.30 0.03 0.10 publ
741 0.25 0.75 0.00 0.00 publ
```

Auflistung 6.11: Erweiterung eines Feature-Sets um Klassenlabel

Für die Klassifikationsaufgaben wird in dieser Arbeit das Klassifikationspaket `caret` (Kuhn u. a. 2018) verwendet, das viele Optionen für Parameter-Tuning über ein einheitliches Interface für verschiedene Klassifikatormodelle zur Verfügung stellt; für die **Random-Forest**-Klassifikation wird die `caret`-Hauptfunktion `train` mit der Methode `rf` aufgerufen (`method=rf`), die die Implementierung des Random-Forest-Algorithmus durch das `randomForest`-Paket (von Liaw & Wiener 2018, basierend auf Breiman 1996; 2002) verwendet.

```
set.seed(1)
train_control <- trainControl(method="cv", number=10)
grid <- data.frame(mtry=sqrt(ncol(featset))
model <- train(GENRE~., data=classdata, method="rf", trControl=train_control,
     tuneGrid=grid)
```

Auflistung 6.12: Training eines Random-Forest-Klassifikators

Zu Beginn des Random-Forest-Modellaufbaus wird zur Reproduzierbarkeit des Ergebnisses vor jedem Aufruf von der `caret`-Hauptfunktion `train`, die das Training des Klassifikators durchführt, ein sog. Seed gesetzt, das sicherstellt, dass bei der Randomisierung immer dieselben Zufallszahlen verwendet werden und damit bei Wiederholung dieselben Ergebnisse errechnet werden.

Mit dem Defaultwert für den `ntree`-Parameter (Anzahl an Bäumen) produziert das `rf`-Paket 500 randomisierte Entscheidungsbäume. Bei der Konstruktion jedes Baumes werden bei jedem Split $m = \sqrt{p}$ zufällig gesampelte Features verwendet (also $\sqrt{p}$ als Standardwert; p = Features/Prädikatoren; s. Liaw & Wiener 2018: 18; vgl. 4.3.2). Dieser für die Random-Forest-Klassifikation zentrale Tuning-Parameter `mtry` des `rf`-Pakets kann in `caret` durch ein Tuning-Grid über verschiedene Methoden (u. a. *random search* oder *random grid*) optimiert werden; da in dieser Arbeit Vergleiche zwischen Modellen (Modellaccuracy bzw. Vergleich der Kappa-Werte) im Vordergrund stehen und nicht die Optimierung von Modellen (vgl. Mitchell 2017: 2), wird auf diese Tuning-Methoden verzichtet und durchgehend der Defaultwert $\sqrt{p}$ verwendet, der über den `tuneGrid`-Parameter an die `train`-Funktion übergeben wird.

Wie in Abschnitt 4.3 beschrieben, basiert der Random-Forest-Klassifikator auf wiederholtem Sampling sowohl der Features pro Split (`mtry`) als auch der Trainingsdaten (doppelte Randomisierung). Dieses Resampling der Trainingsdaten geschieht über **Bootstrapping**, d. h. einer Zufallsauswahl von N Exemplaren aus den N Trainingsdaten, wobei ein Trainingsexemplar mehrfach vorkommen kann, d. h. mit Ersatz/Wiederholung (Defaultwert `rf`-Paket: `replace=TRUE`).

Das Bootstrapping-Resampling der Trainingsdaten dient also der wiederholten Randomisierung der Trainingsdaten für die Verwendung als Ensemble-Klassifikator (Bagging = Bootstrap Aggregation, s. 4.3.2). Daneben wird hier mit der **Kreuzvalidierung** (`cv` = Cross-Validation) zusätzlich eine weitere Methode des Resamplings der Trainingsdaten eingesetzt; diese Methode hält systematisch einen definierten Teil der Trainingsdaten zurück (sog. *folds*) und nutzt diese nicht im Training verwendeten Daten dann zur Bestimmung der Modellaccuracy (s. James u. a. 2017: 176ff.). Cross-Validation wird also als Resampling-Methode zur Abschätzung der Fehlerrate eines Modells ohne explizites Testset verwendet (s. James u. a. 2017: 176).

Gesteuert wird die Resampling-Methode über den `trControl`-Parameter, dem man die `trainControl`-Methode übergibt; mit `cv` werden wiederholt Random-Forest-Modelle, basierend auf gesampelten Trainingsdaten, erstellt und die mittlere Accuracy dieser Modelle berechnet; der `number`-Parameter bestimmt die Anzahl der Wiederholungen (*folds*), die auf den Trainingsdaten erstellt werden. Üblich ist eine – wie hier durchgeführte – *10-fold-cross-validation* (s. James u. a. 2017: 183),[22] die die Trainingsdaten in zehn Partitionen teilt und schrittweise jeweils eine davon auslässt und zur Berechnung der Accuracy verwendet.

Alternativ kann auch direkt aus den im Bootstrap-Sampling ausgelassenen Exemplaren (den sog. *out-of-bootstrap-* bzw. *out-of-bag*-Samples) die Accuracy bzw. Fehlerrate bestimmt werden (Breiman & Cutler 2020; Breiman 2001: 11; vgl. Hastie & Tibshirani & Friedman 2009: 592f.). Wählt man als `trControl`-Methode `oob`, wird ein Random-Forest-Modell erstellt und die Accuracy basierend auf der OOB-Fehlerrate berechnet.

```
importance <- varImp(model, scale=FALSE)
plot(importance)
```

Auflistung 6.13: Berechnung der Feature-Importance

Die **Feature-Importance** als die mittlere Abnahme der Fehlerrate (hier des Gini-Index) kann anschließend auf dem trainierten Random-Forest-Modell berechnet werden (s. 4.3.3, Embedded-Feature-Selection). Das `caret`-Paket stellt dafür die `varImp`-Funktion zur Verfügung; mit `scale=FALSE` wird die originale Skala beibehalten und eignet sich so zur Beurteilung der absoluten wie relativen Relevanz von Merkmalen in der hier angestrebten Feature-Analyse.

6.1.4.2 Klassifikation von Frequent-Patterns-Feature-Sets

Frequent-Patterns-Feature-Sets können grundsätzlich analog zu anderen Feature-Sets mit Random-Forest klassifiziert werden; ggf. bietet sich hier bei aufgrund automatischer Frequent-Pattern-Extraktion hochdimensionalen Feature-Sets eine Feature-Subset-Selection an, d. h. eine Auswahl der relevantesten Features über den Feature-Importance-Rang (s. 4.4.4; vgl. Xing & Pei & Keogh 2010: 41). Aufgrund des hohen `pmin.supports` bei der Pattern-Extraktion (s. 6.1.2.6) ist ein solches Feature-Subsetting in dieser Arbeit nicht notwendig.

6.1.4.3 Sequenz-Klassifikation mit knn (k-Nearest-Neighbour)

Wie in Abschnitt 4.3 erläutert, sind die Feature-basierten Klassifikationsmethoden (hier insbesondere Random-Forest) nicht für sequentielle Textrepräsentationen

22 10- oder 5-folds ergeben einen Kompromiss zwischen Verzerrung und Varianz des Klassifikators („bias-variance trade-off", James u. a. 2017: 184).

geeignet.[23] Stattdessen kommen die zwei in 4.4.3 beschriebenen Methoden für die Klassifikation von Sequenzen zum Einsatz, nämlich **k-Nearest-Neighbour** (knn) und String-Kernel-basierte Support-Vector-Machine-Methoden (**Spectrum-SVM**).

Die knn-Sequenzklassifikation verwendet eine Sequenz-Distanzmatrix (s. 6.1.3.2; vgl. Xing & Pei & Keogh 2010), auf der basierend (zu gegebenem k) die k-Nachbarn ermittelt werden (die Datenpunkte, die nach der gegebenen Distanzmatrix am nächsten liegen); dazu wird hier die `k.nearest.neighbors`-Funktion von `fastKNN` verwendet.[24] Für die verschiedenen Textsortenkategorisierungen wird anschließend aus den Daten der k-Nachbarn pro Text jeweils die Klasse errechnet, in der die meisten der Nachbarn des Textes liegen; diese Klasse wird als Vorhersage verwendet und die Accuracy im Vergleich mit der tatsächlichen Klasse jedes Textes berechnet.[25]

6.1.4.4 Sequenzklassifikation mit Spectrum-SVM

Eine Alternative zur direkten Klassifikation von Sequenzen mit dem knn-Algorithmus ist die Klassifikation mit Support-Vector-Machine-Methoden (SVM) mit String-Kernel (**k-Spectrum-String-Kernel**), die eine Sequenz in einen Merkmalsvektor aus Teilsequenzen der Länge k aus dem Alphabet der Sequenz (Spektrum) überführt und damit in einen höherdimensionalen Raum transformiert (s. 4.4.3; Xing & Pei & Keogh 2010). Der im Folgenden Spectrum-SVM genannte Klassifikator (im `caret`-Paket als `svmSpectrumString`-Methode des `kernlab`-Pakets[26] implementiert) wird parallel zu den Sequenzklassifikationen mit knn erprobt.

Für diesen Typ der Sequenzenklassifikation wird die Liste der Sequenzfolgen unter Verwendung eines Alphabets mit Zeichen der Länge 1 recodiert (z. B. wird die Folge `SWITCH-CONT-CONT` zu `SCC`); diese Klassenlabels werden dann der `train`-Funktion von `caret` als Liste übergeben.

```
set.seed(1)
train_control <- trainControl(method = "cv", number = 3, returnResamp = "all")
model <- train(svm_data, base, method = "svmSpectrumString", trControl =
     train_control)
```

Auflistung 6.14: Sequenzklassifikation mit Spectrum-SVM

23 Entsprechend entfällt hier auch eine Feature-Importance-Bestimmung.

24 Als Distanzmatrix für eine solche *sequence distance based classification* (Xing & Pei & Keogh 2010) können sowohl die DTW- als auch Edit-Distance-Abstände verwendet werden.

25 Die knn-Klassifikation wurde im Rahmen dieser Fallstudie nur testweise im Sinne einer Methodenexploration ohne Cross-Validation durchgeführt.

26 Dokumentation unter https://cran.r-project.org/package=kernlab (abgerufen am 03.09.2022).

6.1.5 Vorgehen zur Feature-Projection

PCA-Clusterplots. Die Hauptkomponentenanalyse (PCA) wird mit der `clusplot`-Funktion des `cluster`-Pakets[27] für die dimensionsreduzierte Darstellung von Clustertypologien durchgeführt. Durch Übergabe der Datenmatrix des Feature-Sets und der zuvor gefundenen Clustergruppierung werden die beiden ersten Hauptkomponenten aufgetragen und die Trennung der Clustergruppen kann so beurteilt werden.

```
library(cluster)
clusplot(cluster_data, groups, diss = FALSE)
```

Auflistung 6.15: Berechnung PCA-Clusterplot

Visualisierung von Sequenzdistributionen. Für eine Visualisierung von Sequenzen bietet das `TraMineR`-Paket verschiedene Methoden wie Sequenzdistributionen (Chronogramme), gruppierte Sequenz-Indexplots oder Durchschnittszeit-Plots an (s. 4.4.5); die einzelnen Typen können über die `seqplot`-Funktion des `TraMineR`-Pakets ausgewählt werden (vgl. Gabadinho u. a. 2011).

Barycenter-Kurven von DTW-alignierten Sequenzen. Zur Darstellung von Gruppen von DTW-alignierten Sequenzen wird mit der `DBA`-Funktion des `dtwclust`-Pakets (Sarda-Espinosa 2018) pro Gruppe eine Barycenter-Kurve von deren Sequenzen berechnet und diese – zusammen mit den einzelnen Sequenzen, deren Durchschnitt diese Kurve darstellt – geplottet (vgl. 4.4.5; Petitjean & Ketterlin & Gançarski 2011).

6.2 Globale morphosyntaktische Textstruktur-Typologie

Für die globalen morphosyntaktischen Textstrukturmodelle wird je nach Strukturebene ein Primärdatensatz aus der Datenbank erzeugt. Dabei werden bereits hier in einer vorläufigen Feature-Construction einfache Frequenzdaten berechnet (z. B. die Phrasen pro Clause = `PHRASE_COUNT`). Anschließend werden die textweiten kognitiven Texttypologie-Parameter berechnet, zu Feature-Sets zusammengefasst und ausgewertet. Wie oben ausgeführt, müssen die zu einem Feature-Set zusammengefügten textweiten (globalen) Merkmale aufgrund ihrer disparaten Skalen zur Vergleichbarkeit der Feature-Dimensionen skaliert werden. Für folgende Strukturdimensionen werden jeweils Datensätze generiert:

27 Dokumentation unter https://cran.r-project.org/web/packages/cluster/cluster.pdf#page=23 (abgerufen am 03.09.2022).

- Segmentierungsebene **Token**
- Segmentierungsebene **Phrase**
- Segmentierungsebene **Clause**
- Segmentierungsebene **Satz**

	SEGMENTS	FUNCTIONS	SENT_NR	FORM	LEMMA	Text-ID
1	0	0	1	teːm qatəɬ	teːm qatəɬ	728
2	0	0	1	mɐː	mɐː	728
3	0	0	1	βɐːr	βɐːr	728
4	1	1	1	ɬɒːtʲtʲɐ	ɬɒːtʲ	728
5	1	1	1	jaŋqəm	jaŋq	728
6	0	0	2	βɐːr	βɐːr	728

Report 6.2.1: Token-Datensatz (Globale Parameter)

	TOKEN_COUNT	PHR	SYN	SEM	Text-ID
1	1	advP	ADV		728
2	1	pronP	S		728
3	1	attrP	ATTR	INANIM	728
4	1	compC	SUBPRED	PERCEPTION	728
5	1	finVP	PRED	MOTION	728
6	0	zero	S		728

Report 6.2.2: Phrasen-Datensatz (Globale Parameter)

	PHRASE_COUNT	SUB_CLAUSE_COUNT	CLAUSE_TYPE	Text-ID
1	5	1	finVP	728
2	8	2	finVP	728
3	1	0	finVP	728
4	1	0	finVP	728
5	3	0	finVP	728
6	4	0	finVP	728

Report 6.2.3: Clause-Datensatz (Globale Parameter)

	SUB_CLAUSE_COUNT	SENT_PHR_LENGTH	CLAUSE_TYPE	Text-ID
1	1	5	finVP	728
2	2	10	finVP	728
3	0	8	finVP	728
4	0	10	passVP	728
5	0	17	PTCL	728
6	0	9	finVP	728

Report 6.2.4: Satz-Datensatz (Globale Parameter)

Im Folgenden werden zunächst einzelne Teilmodelle für verschiedene textstrukturelle Dimensionen untersucht (allgemeines, referentielles sowie relationales Modell), bevor die Auswertung des Gesamtmodells aller Merkmale erfolgt.

6.2.1 Globales Grundparameter-Modell

- **Beobachtungsgegenstand**: globale textstrukturelle Merkmale
- **Feature-Attribute**: morphosyntaktische Text-Eigenschaften
- **Feature-Werte**: Clause-Elaboration; Komplexität; Redundanz; lexikalische Dichte
- **Normierung der Features**: relative Texthäufigkeiten bzw. Guiraud-Index
- **Skalierung des Feature-Sets**: z-Standardisierung

Feature-Construction und -Extraction. Folgende Werte werden pro Text aus den im Datensatz vorhandenen Frequenzwerten berechnet:

- **Clause-Elaboration** als Durchschnitt der Phrasenanzahl pro Clause (Clause-Datensatz)
- **Clause-** bzw. **Satz-Komplexität** als Durchschnitt der Anzahl an subordinierten Verbalkonstruktionen pro Clause bzw. Satz (Clause-/Satz-Datensatz)
- **Redundanz** als Verhältnis Tokenanzahl zu Typeanzahl (Token-Datensatz)
- **Lexikalische Dichte** als Verhältnis Lemma-Types zur Wurzel der Tokenanzahl, d. h. normiert über Guiraud-Index (Token-Datensatz)

Folgendes Feature-Set ergibt sich:

Text-ID	CL_ELAB	CL_COMPLEX	SENT_COMPLEX	RED	LEX_DENS
728	3.27	0.15	0.31	1.45	6.37
730	3.04	0.18	0.33	1.54	7.79
732	2.94	0.19	0.33	1.72	7.81
741	2.25	0.00	0.00	1.10	3.02
742	2.63	0.08	0.18	1.46	8.21
750	2.50	0.08	0.19	2.34	8.54

Report 6.2.5: Unskaliertes Feature-Set (Globale Grundparameter)

Da die globalen Textstruktur-Merkmale unterschiedliche Skalen besitzen, erfolgt anschließend eine Skalierung der Merkmale des Feature-Sets: Durch z-Standardisierung wird für jedes Feature eine Zentrierung seines Mittelwerts auf 0 und eine Skalierung gemäß seiner Streuung vorgenommen (s. 4.1.3.2). Die Unterschiede von Werten innerhalb einer Feature-Dimension in Report 6.2.6 sowie in den folgenden gruppenbezogenen Analysen zeigen durch diese Skalierung die positive oder negative Abweichung von dem auf 0 zentrierten Mittelwert an.

Nach der z-Standardisierung stellt sich das Feature-Set folgendermaßen dar:

Text-ID	CL_ELAB	CL_COMPLEX	SENT_COMPLEX	RED	LEX_DENS
728	0.33	−0.04	0.43	−0.58	0.06
730	−0.03	0.19	0.55	−0.34	0.83
732	−0.18	0.27	0.55	0.11	0.84
741	−1.23	−1.30	−1.44	−1.50	−1.77
742	−0.65	−0.65	−0.35	−0.56	1.06
750	−0.85	−0.65	−0.32	1.75	1.24

Report 6.2.6: Skaliertes Feature-Set (Globale Grundparameter)

Im Folgenden sind die zur Extraktion der Features in R erfolgten aggregierenden Berechnungen von Mittelwerten und anderen auf textweite Werte reduzierenden Operationen unter Verwendung der basalen Merkmale in den Primärdatensätzen (s. Reports 6.2.1–6.2.4) aufgelistet:

```
CL_ELAB = mean(x$PHRASE_COUNT)
```

Auflistung 6.16: Feature-Construction für Clause-Elaboration: Phrasen pro Clause (Globale Grundparameter)

```
CL_COMPLEX = mean(x$SUB_CLAUSE_COUNT)
```

Auflistung 6.17: Feature-Construction für Clause-Komplexität: Subordinierte Elemente pro Clause (Globale Grundparameter)

```
SENT_COMPLEX = mean(x$SUB_CLAUSE_COUNT)
```

Auflistung 6.18: Feature-Construction für Satz-Komplexität: Subordinierte Elemente pro Satz (Globale Grundparameter)

```
RED = length(x$FORM) / length(unique(x$FORM))
```

Auflistung 6.19: Feature-Construction für Redundanz: Token-Type-Verhältnis (Globale Grundparameter)

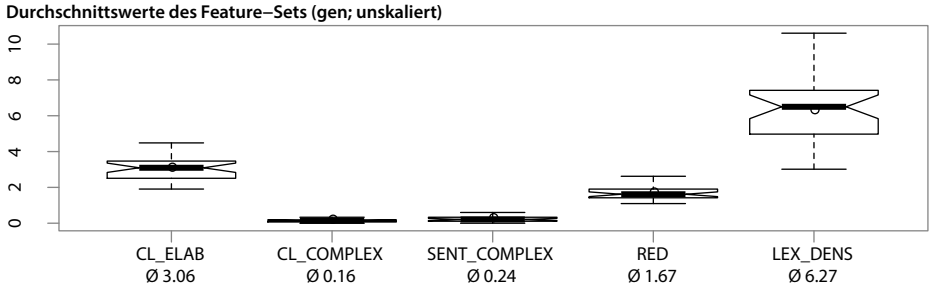
```
LEX_DENS = length(unique(x$LEMMA)) / sqrt(length(x$FORM))
```

Auflistung 6.20: Feature-Construction für lexikalische Dichte (Globale Grundparameter)

Ergebnisse. Eine Auswertung der unskalierten Durchschnittswerte der allgemein-globalen Features der untersuchten obugrischen Texte zeigt (s. Plot 6.2.1):

- eine durchschnittliche **Clause-Elaboration** von drei Phrasen pro Satz
- eine durchschnittliche **Clause-Komplexität** von 0.16 `SUBPRED`s/Clause (1 `SUBPRED` auf 6 Clauses)

- eine durchschnittliche **Satz-Komplexität** von 0.24 SUBPREDs/Satz (1 SUBPRED auf 4 Sätze)
- eine durchschnittliche **Redundanz** von 1.67 Tokens/Types
- eine durchschnittliche **lexikalische Dichte** von 6.27 Lemma-Types/√Token



Plot 6.2.1: Boxplot des unskalierten Feature-Sets (Globale Grundparameter)

Im Vergleich der absoluten Werte zeigt sich demnach u. a. eine geringe Komplexität (etwa im Vergleich mit den Komplexitätswerten udischer Volksmärchen in Schulze 2004a: 555), d. h. es werden in den obugrischen Volkserzählungen verhältnismäßig wenige Hintergrundinformationen gegeben.

Für den weiteren Vergleich der Feature-Werte untereinander und bzgl. der Differenzen in den Apriori-Kategorisierungsgruppen werden im Folgenden die skalierten Werte herangezogen (d. h. zentriert auf Mittelwert = 0).

Die stärkste Abweichung vom Durchschnitt des Korpus findet sich bei den Zeitungstexten (`jou` in Plot A.4), die im Gegensatz zu den Volkserzählungen gekennzeichnet sind durch hohe Komplexitätswerte (d. h. viele Hintergrundinformationen), starke Clause-Elaboration (d. h. hohe Informationsdichte), geringe Redundanz und hohe lexikalische Dichte, was auf ein TWM mit vielen neuen, aber sich nur schwach wiederholenden Konzepten hindeutet.

Die narrativen Texte (`narr` in Plot A.7) haben u. a. eine höhere Redundanz, was mit der erwartbaren Prototypik für Volkserzählungen korrespondiert und als Hinweis auf den „Aufbau eines geschlossenen referentiellen Wissens“ (Schulze 2020: 607) gesehen werden kann.

Clustering (Plot A.1). Folgende Typen lassen sich in der gewählten Clustertypologie mit drei Gruppen unterscheiden (s. Plots A.3f.):

- Die periphere Gruppe (3) mit einem `fas`-, einem `jou`- und einem `tal`-Text zeichnet sich durch relative hohe Komplexitätswerte aus.
- Der Hauptcluster unterteilt sich in eine Subgruppe (2), die primär die längeren Erzähltexte enthält (Zaubermärchen = `magic_tal`), die wenig elaboriert und

redundant sind (mit höherer Redundanz, niedrigerer Komplexität und Clause-Elaboration, also niedrigerem Informationsgehalt pro Clause, als der Durchschnitt), und
- in die Gruppe 1 als Kerngruppe, deren Werte nahe am Durchschnitt der globalen Grundparameter-Werte liegen. Sie enthält vor allem `priv`-Texte (gemäß der `COMM_SIT`-Kategorisierung, also Texte aus Interview-artiger Kommunikationssituation), das sind vor allem die kurzen Fabel-Tiermärchen (`anim_tal`) und ethnographischen Berichte (`eth`), die sich im Vergleich mit den klassischen Märchen vor allem durch eine höhere Informationsdichte (= höhere `CL_ELAB`; mehr Informationen per Clause als Informationsgrundeinheit) auszeichnen (vgl. Plot A.5).

Klassifikation. Das globale Grundparameter-Feature-Set erreicht mit 0.74 für die `COMM_SIT`-Kategorisierung einen Kappa-Wert > 0.6 (*substantial agreement*; vgl. 4.3.3.3). Die restlichen Kategorisierungen bleiben unter der Schwelle von 0.4 (*moderate agreement*). Mit Abstand wichtigstes Feature ist in der `COMM_SIT`-Klassifizierung die Clause-Elaboration. Für die anderen Klassifizierungen ist die lexikalische Dichte (`BASE`, `GENRE`) bzw. die Redundanz (`DISC_STRUCT`) wichtigstes Merkmal, die Clause-Elaboration ist jeweils das zweitwichtigste (s. Plots A.8ff.).

6.2.2 Globales nominal-referentielles Modell

- **Beobachtungsgegenstand**: globale Textstruktur
- **Feature-Attribute**: morphosyntaktische Text-Eigenschaften
- **Feature-Werte**: referentielle Explizitheit, referentielle Inferenz, nominale Elaboration, referentielle Dichte
- **Normierung der Features**: Textdurchschnitt bzw. relative Textfrequenz
- **Skalierung des Feature-Sets**: z-Standardisierung

Feature-Construction und -Extraction. Folgende Werte werden pro Text aus den im primären Phrasen-Datensatz vorhandenen Frequenzwerten berechnet:

- **Referentielle Inferenz** als Verhältnis der Anzahl nicht-overter referentieller Einheiten (zero-Anaphern und Possessivsuffixe) zu Gesamtzahl referentieller Einheiten
- **Pronominale Inferenz** als Verhältnis der Anzahl pronominaler referentieller Einheiten (partiell-covert) zu Gesamtzahl referentieller Einheiten
- **Referentielle Explizitheit** als Verhältnis Tokenanzahl lexikalischer nominaler Einheiten zu Gesamtzahl substantieller referentieller Einheiten
- **Nominale Elaboration** als durchschnittliche Tokenanzahl pro lexikalischer nominaler Einheit
- **Referentielle Dichte** als Verhältnis der Anzahl von overt kodierten Argument-Phrasen zu Anzahl von Argument-Positionen (vgl. Bickel 2003: 726)

Die Normierung bzgl. der unterschiedlichen Textlänge geschieht über die Berechnung von Durchschnittswerten und Verhältniswerten. Folgendes Feature-Set ergibt sich:

Text-ID	REF_INFER	PRON_INFER	REF_EXPLIC	NOM_ELAB	REF_DENS
728	0.42	0.09	1.24	1.54	0.50
730	0.41	0.08	1.40	1.68	0.48
732	0.36	0.15	1.14	1.49	0.66
741	0.44	0.00	1.40	1.40	0.43
742	0.48	0.01	1.42	1.63	0.46
750	0.53	0.17	1.27	2.04	0.51

Report 6.2.7: Unskaliertes Feature-Set (Global-referentielle Parameter)

Nach der Skalierung stellt sich das Feature-Set folgendermaßen dar:

Text-ID	REF_INFER	PRON_INFER	REF_EXPLIC	NOM_ELAB	REF_DENS
728	0.26	0.26	−0.61	−0.30	−0.83
730	0.17	0.01	−0.09	0.25	−0.96
732	−0.23	1.14	−0.96	−0.47	0.33
741	0.44	−1.19	−0.08	−0.82	−1.35
742	0.75	−1.02	0.00	0.05	−1.11
750	1.11	1.52	−0.53	1.58	−0.74

Report 6.2.8: Skaliertes Feature-Set (Global-referentielle Parameter)

Im Folgenden sind die zur Extraktion der Features in R erfolgten aggregierenden Berechnungen von Mittelwerten und anderen auf textweite Werte reduzierenden Operationen aufgelistet:

```
REF_INFER = length(x$REF[x$REF!=0 & (x$PHR=='zero' | x$PHR=='px') ]) /
     length(x$REF[x$REF!=0])
```

Auflistung 6.21: Feature-Construction für referentielle Inferenz (Global-referentielle Parameter)

```
PRON_INFER = length(x$REF[x$REF!=0 & x$PHR=='pronP' ]) / length(x$REF[x$REF!=0])
```

Auflistung 6.22: Feature-Construction für pronominale Inferenz (Global-referentielle Parameter)

```
REF_EXPLIC = sum(x$TOKEN_COUNT[x$REF!=0 & x$TOKEN_COUNT!=0 & x$PHR!="pronP"]) /
     length(x$REF[x$REF!=0 & x$TOKEN_COUNT!=0])
```

Auflistung 6.23: Feature-Construction für referentielle Explizitheit (Global-referentielle Parameter)

```
NOM_ELAB = mean(x$TOKEN_COUNT[x$PHR=='NP' | x$PHR=='locNP' | x$PHR=='postP'])
```

Auflistung 6.24: Feature-Construction für nominale Elaboration (Global-referentielle Parameter)

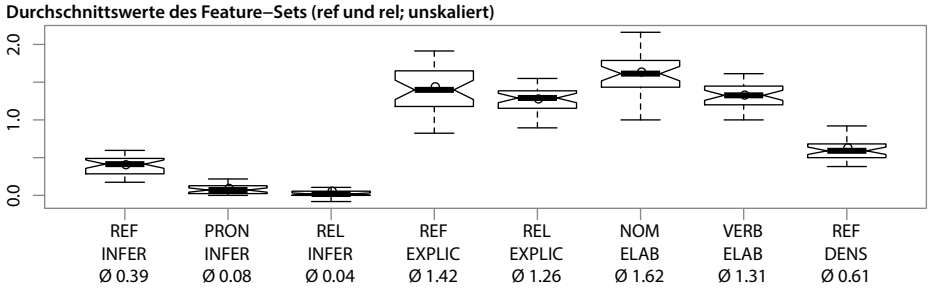
```
REF_DENS = length(x$SYN[(x$SYN=='S' | x$SYN=='O') & x$PHR!='zero']) /
     length(x$SYN[x$SYN=='S' | x$SYN=='O'])
```

Auflistung 6.25: Feature-Construction für referentielle Dichte (Global-referentielle Parameter)

Ergebnisse. Eine Auswertung der unskalierten Durchschnittswerte der referentiellen globalen Features der untersuchten obugrischen Texte zeigt (s. Plot 6.2.2):

- 39% aller Referentenerwähnungen sind nicht-overt (hohe absolute referentielle Inferenz, auch bedingt durch pro-drop).
- Das korrespondiert mit 61% overten Argument-Phrasen (= referentielle Dichte als Form der invertierten referentiellen Inferenz, bezogen auf Argumentstellen).
- 8% der Referentenerwähnungen sind partiell-covert (pronominale Inferenz).

Die restlichen absoluten Werte der globalen nominal-referentiellen Features werden im Folgenden im Vergleich mit den relationalen besprochen.



Plot 6.2.2: Boxplot des unskalierten Feature-Sets (Global-referentielle und global-relationale Parameter)

Die gruppenspezifische Auswertung der skalierten Werte (s. Plots B.5ff.) zeigt für die längeren Zaubermärchen (`magic_tal`) und die Mythen – im Gegensatz vor allem zu den Zeitungstexten – eine **erhöhte referentielle Inferenz** (und entsprechend eine niedrige referentielle Dichte, also viele Nullanaphern) bei gleichzeitig **niedriger referentieller Explizitheit** (im Gegensatz zu den Fate Songs). Im Gegensatz dazu haben Fabel-Tiermärchen (`anim_tal`) eine geringere Inferenzstärke (höhere referentielle

Dichte).[28] Denkbar wäre hier ein Einfluss russischer Erzähltradition[29] auf die Entstehung des Subgenres obugrischer Zaubermärchen, wonach die kurzen Tierfabeln ein älteres narratives TWM darstellen würden; vgl. in dem Zusammenhang auch Schulze (2004a: 556, 573) zur Entwicklung udischer Volkserzählungen hin zu mehr referentieller Inferenz und weniger Explizitheit unter einer Anpassung an ein westliches narratives TWM, wodurch diese ‚relationaler' werden („[*Modern*] Nizh speakers seem to rely more strongly on interactional features", Schulze 2004a: 573; s. auch 6.2.3: relationale Parameter; 6.5.6: Dialogstrukturen).

Clustering (Plot B.1). Im Clustering zeigen sich zwei Gruppen (s. Plot B.3):

- Gruppe 2 enthält vor allem nicht-narrative Texte und einige der kurzen Fabel-Texte (insbesondere beide `BeardedMan`-Varianten) mit:
  - weniger referentieller Inferenz als der Durchschnitt; höherer referentieller Dichte
  - höherer referentieller Explizitheit und Elaboration

28 Vgl. folgenden Ausschnitt aus dem fabelartigen nordmansischen Märchen 1232 (`BeardedMan`, `anim_tal`) mit hoher referentieller Dichte:

(3) tus-əŋ ojka ulʲa ta peːlamt-as
moustache-PROPR old_man fire EMPH1 ignite-PST[3SG]
Old Man with a Beard lit a fire. (NM, 1232: 5)

(4) ula jomasʲ-akʷe at pelamt-i
fire good-DIM NEG ignite-PRS[3SG]
The fire doesn't burn well. (NM, 1232: 6)

(5) tus-əŋ ojka ulʲa ta puw-l-i-te
moustache-PROPR old_man fire EMPH1 blow-DUR-PRS-SG<3SG
Old Man with a Beard blows on the fire. (NM, 1232: 7)

Vgl. dagegen einen Ausschnitt aus dem Zaubermärchen 1237 (`ThreeSons`, `magic_tal`) mit niedriger referentieller Dichte:

(6) oːwlət neː-nəl palt xoːj-as-ət
at_the_beginning woman-ABL to encounter-PST-3PL
First they came to the woman. (NM, 1237: 108)

(7) poːxan keːt-s-anəl
away send-PST-SG<3PL
They sent her off. (NM, 1237: 109)

(8) ta poːxan min-as
EMPH1 away go-PST[3SG]
She went off. (NM, 1237: 110)

(9) xoːntl-aŋkʷe tuwəl pat-s-ət
to_be_at_war-INF then begin-PST-3PL
Then they started to fight. (NM, 1237: 111)

29 Vgl. Propps auf russischen Erzählungen basierende Zaubermärchen-Typologie (1972).

- Gruppe 1 enthält vor allem narrative Texte (insbesondere Zaubermärchen sowie die Mythen) mit:
  - höherer referentieller Inferenz (niedrigerer referentieller Dichte)
  - geringerer referentieller Explizitheit und Elaboration

Klassifikation. Für das globale nominal-referentielle Feature-Set erreicht keine Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*).

6.2.3 Globales verbal-relationales Modell

- **Beobachtungsgegenstand**: globale Textstruktur
- **Feature-Attribute**: morphosyntaktische Text-Eigenschaften
- **Feature-Werte**: relationale Explizitheit, relationale Inferenz, verbale Elaboration
- **Normierung der Features**: Textdurchschnitt bzw. relative Textfrequenz
- **Skalierung des Feature-Sets**: z-Standardisierung

Feature-Construction und -Extraction. Folgende Werte werden pro Text aus den im Phrasen-Datensatz vorhandenen Frequenzwerten berechnet:

- **Relationale Inferenz** als Verhältnis der Anzahl non-verbaler relationaler Einheiten (als indirekter Hinweis auf Verbellipsen im Korpus) zu Gesamtzahl relationaler Einheiten[30]
- **Relationale Explizitheit** als Verhältnis Tokenanzahl verbaler Einheiten zu Gesamtzahl relationaler Einheiten[31]
- **Verbale Elaboration** als durchschnittliche Tokenanzahl pro lexikalischer verbaler Einheit

Die Normierung bzgl. der unterschiedlichen Textlänge geschieht über die Berechnung von Durchschnittswerten und Verhältniswerten. Report 6.2.9 zeigt das Feature-Set vor der Skalierung, Report 6.2.10 danach.

30 Für die relationale Inferenz wird der Anteil von Verbellipsen bestimmt über die im Korpus als Prädikat (`PRED` bzw. `SUBPRED`) ausgezeichneten non-verbalen Einheiten (in der Annotation ist in jeder Clause-artigen syntaktischen Einheit ein Element als `PRED` bzw. `SUBPRED` ausgezeichnet, d. h. in Nominalsätzen ggf. das Subjekt).

31 Für die relationale Explizitheit werden die in prädikativer Funktion verwendeten non-verbalen Einheiten als relationale Proformen aufgefasst, also den Pronomina in der Operationalisierung von referentieller Explizitheit als substantielle, aber nicht-explizite Formen entsprechend; dies ist eine andere Interpretation dieser Formen als bei der relationalen Inferenz, wo sie als Kennzeichen der Verbellipse dienen.

Text-ID	REL_INFER	REL_EXPLIC	VERB_ELAB
728	0.00	1.23	1.23
730	0.11	1.35	1.51
732	0.10	1.30	1.45
741	0.00	1.00	1.00
742	0.04	1.55	1.61
750	0.04	1.36	1.42

Report 6.2.9: Unskaliertes Feature-Set (Global-relationale Parameter)

Text-ID	REL_INFER	REL_EXPLIC	VERB_ELAB
728	−0.46	−0.18	−0.54
730	0.78	0.48	1.26
732	0.73	0.24	0.89
741	−0.46	−1.45	−2.01
742	0.00	1.60	1.92
750	0.05	0.54	0.69

Report 6.2.10: Skaliertes Feature-Set (Global-relationale Parameter)

Im Folgenden sind die zur Extraktion der Features in R erfolgten aggregierenden Berechnungen von Mittelwerten und anderen auf textweite Werte reduzierenden Operationen aufgelistet:

```
REL_INFER = ( length(x$PHR[x$SYN=="PRED" | x$SYN=="SUBPRED"]) -
     length(x$PHR[x$PHR=='finVP' | x$PHR=='passVP' | x$PHR=='okVP' |
     x$PHR=='ptcpVP' | x$PHR=='infVP' | x$PHR=='subC' | x$PHR=='compC' |
     x$PHR=='CVB']) )
/
length(x$PHR[x$SYN=="PRED" | x$SYN=="SUBPRED"])
```

Auflistung 6.26: Feature-Construction für relationale Inferenz (Global-relationale Parameter)

```
REL_EXPLIC = sum(x$TOKEN_COUNT[x$PHR=='finVP' | x$PHR=='passVP' | x$PHR=='okVP'
     | x$PHR=='ptcpVP' | x$PHR=='infVP' | x$PHR=='subC' | x$PHR=='compC' |
     x$PHR=='CVB'])
/
length(x$SYN[x$SYN=='PRED' | x$SYN=='SUBPRED'])
```

Auflistung 6.27: Feature-Construction für relationale Explizitheit (Global-relationale Parameter)

```
VERB_ELAB = mean(x$TOKEN_COUNT[x$PHR=='finVP' | x$PHR=='passVP' | x$PHR=='okVP'
     | x$PHR=='ptcpVP' | x$PHR=='infVP' | x$PHR=='subC' | x$PHR=='compC' |
     x$PHR=='CVB'])
```

Auflistung 6.28: Feature-Construction für verbale Elaboration (Global-relationale Parameter)

Ergebnisse. Eine Auswertung der unskalierten Durchschnittswerte der relationalen globalen Features (vgl. Plot 6.2.2) der untersuchten obugrischen Texte im Vergleich mit den referentiellen Werten zeigt:

- Es gibt bedeutend mehr referentielle (0.39) als relationale Inferenz (0.04) (dies ist wegen pro-drop erwartbar).

- Die Stärke der referentiellen und relationalen Explizitheit ist relativ ähnlich (1.3 bzw 1.4 Tokens pro Einheit);
- ebenso die Stärke der Elaboration.

Die gruppenspezifische Auswertung der skalierten Werte (s. Plots C.6ff.) zeigt insbesondere für die narrativen Texte eine **relativ niedrige relationale Inferenz** (im Gegensatz vor allem zu den `eth`- und `fas`-Texten, `jou` dagegen ähnlich) bei gleichzeitig **höherer relationaler Explizitheit** (im Gegensatz wieder zu den `eth`- und `fas`-Texten).[32] Dies kann man dahingehend interpretieren, dass narrative Texte eher handlungsorientiert sind und entsprechend eine hohe relationale Explizitheit an den Tag legen sowie eine niedrige relationale Inferenz, während die autobiographischen Fate Songs als Personal Songs eher mit Orten, Dingen und Menschen verknüpfte Empfindungen vermitteln (vgl. Ojamaa & Ross 2004: 134) und die journalistischen Texte über Menschen, Orte und Dinge berichten (also jeweils referentiell expliziter sind).

Clustering (Plot C.1). Im Clustering weist das globale relationale Feature-Set den besten Cluster-Tendency-Wert für alle globalen Feature-Sets auf (0.25 Hopkins); es zeigt sich neben den relativ ähnlichen Hauptgruppen (1) und (2) eine kleine Cluster-Gruppe mit den beiden nicht-narrativen Texten 1355 (`eth`) und 1373 (`fas`). Diese Gruppe (3) (s. Plots C.2ff.) hat:[33]

- **höhere relationale Inferenz** (als der Durchschnitt)
- **niedrigere relationale Explizitheit** und Elaboration (als der Durchschnitt)

Klassifikation. Für das globale verbal-relationale Feature-Set erreicht keine Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*).

32 Aber auch innerhalb der narrativen Texte haben die Zaubermärchen (`magic_tal`) eine höhere relationale Explizitheit und Elaboration gegenüber den Fabelmärchen; auch dies spricht für einen ‚relationaleren' narrativen TWM-Subtyp der Zaubermärchen (s. 6.2.2).

33 Folgende Textausschnitte in Übersetzung verdeutlichen diese nicht-narrative Typik mit hoher *relationaler* Inferenz und ebenso hoher *referentieller* Explizitheit und Elaboration; s. dazu auch das Clusteringresultat Plot B.1 der globalen referentiellen Merkmale: Beide Texte sind auch hier zusammen geclustert, allerdings als Teil der größeren Gruppe (2) mit entsprechenden Werten:

- `Trap` = 1355 (`eth`), Satz 5+6: „When you get trough the marsh with the small, short pine trees, there'll be a big bow-trap. That bow-trap... a fallen tree with its roots pulled out... behind that fallen tree with its roots pulled out – go left."
- `HazyDaySong` = 1373 (`fas`), Satz 3+4: „If one looks to the square of the village with a green grass-turfed square, green grass (like) a silken robe stretches to the village square."

6.2.4 Globales Gesamtmodell

- **Beobachtungsgegenstand**: globale Textstruktur
- **Feature-Attribute**: morphosyntaktische Text-Eigenschaften
- **Feature-Werte**: s. Teilmodelle oben
- **Normierung der Features**: s. Teilmodelle oben
- **Skalierung des Feature-Sets**: z-Standardisierung

Die verschiedenen Mengen der bisher behandelten globalen Features (allgemein, referentiell und relational) werden hier zu einem globalen Gesamt-Feature-Set kombiniert und gemeinsam in einer Feature-Evaluation ausgewertet.

Feature-Construction und -Extraction. Das globale Gesamt-Feature-Set basiert auf den bereits besprochenen, skalierten Merkmalen der Einzelmodelle.

Ergebnisse. Für die Ergebnisse der deskriptiven Auswertung siehe die Daten der Einzelmodelle.

Clustering (Plot D.1). Ähnlich wie in den Einzelmodellen zeigt sich im Clustering eine Abtrennung eines Kerns primär narrativer Texte (Cluster 1) von einer Gruppe mit peripheren Genres (`fas`, `jou`, `eth`; Cluster 2). Zusammenfassend zeichnet sich dieser narrative Kern dabei kontrastiv durch folgende Feature-Wert-Tendenzen im referentiellen bzw. relationalen Bereich aus (s. Plot D.2):

- **Inferenz**: [+ referentiell], [- relational]
- **referentielle Dichte**: [-]
- **Explizitheit**: [- referentiell], [+ relational]
- **Elaboration**: [- referentiell], [+ relational]
- **Clause-Elaboration** sowie **Clause-** und **Satz-Komplexität**: [-]
- **Redundanz** und **Lexikalische Dichte**: [+]

Klassifikation. Das globale Gesamt-Feature-Set erreicht für die `COMM_SIT`- und die `DISC_STRUCT`-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle. Der Vergleich der Feature-Importance-Auswertungen der Random-Forest-Klassifikation (s. Plots D.7ff.) für die vier Textsortenklassifizierungen zeigt:

- Die Clause-Elaboration ist wichtigstes Merkmal für die `COMM_SIT`-Klasse.
- Für die `DISC_STRUCT`-Klassifizierung ist dagegen die referentielle Explizitheit wichtigstes Merkmal.

- Für die GENRE-Klassifizierung sind die referentielle bzw. lexikalische Dichte wichtigste Merkmale; ebenso sind die referentielle bzw. pronominale Inferenz relevant.
- Die BASE-Klassifizierung hat kein herausragendes wichtigstes Feature.

6.3 Referenz-Typologie

Die Berechnung der referentiellen TWM-Strukturparameter basiert auf den in einer halbautomatischen Annotation gewonnenen Referenten-Daten des syntaktisch-pragmatisch annotierten Korpus (s. die Referenten-IDs in der REF-Spalte von Report 6.3.1).

	REF	SEM	SYN	PRA	PHR	Text-ID
1	[1]	AG	S	FRAME	pronP	728
2	[5]	PAT	ATTR	FRAME	attrP	728
3	[1]	AG	S	TOP	zero	728
4	[5]	PAT	ATTR	REPEAT	attrP	728
5	[1]	AG	AGR	TOP	px	728
6	[2]	LOC	ADV	FOC	locNP	728

Report 6.3.1: Referentieller Primärdatensatz (Referentielle Parameter)

Im Folgenden werden zunächst die beiden syntagmatischen Topikalitätsparameter der **referentiellen Distanz** sowie der **Topik-Persistenz** als Operationalisierungen von rückwärts- bzw. vorwärtsgewandter Topikalität ausgewertet, bevor deren Merkmale als kombiniertes Gesamtmodell anaphorischer und kataphorischer Referentenstruktur einer Gesamtauswertung unterzogen werden. Anschließend erfolgt die Auswertung des textweiten **Topikalitätsquotienten** (als inverses Maß absoluter referentieller Informationsdichte) sowie eines Feature-Sets der zehn **topikalitätsstärksten Referenten**.

6.3.1 Referentielle Distanz

- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: grammatische Relationen
- **Feature-Wert**: referentielle Distanz
- **Normierung der Features**: Textdurchschnitt pro Feature (GR)
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)
- **Ersatz NAs**: 50 (Defaultwert für unbekannte Referenten)[34]

34 Bei der Abschätzung für fehlenden Werte im Rahmen der Feature-Extraction mit R wurde mit 50 ein Wert verwendet, der leicht von dem in der Feature-Construction verwendeten Maximalwert für den Referentenabstand (60) abweicht. Die Werte liegen aber so nahe beieinander, dass man davon ausgehen kann, dass diese Abweichung keinen relevanten Einfluss auf das Ergebnis hat, zumal da es hier nur um den seltenen Ausnahmefall von Texten geht, in denen eine der grammatischen Relationen nicht vertreten ist.

Feature-Construction. Der Parameter der referentiellen Distanz wird, basierend auf dem Referenten-Datensatz (Report 6.3.1), als der Abstand eines Referenten zu seiner letzten Erwähnung relativ zu dessen aktueller Textposition bestimmt; dabei wird in dieser Arbeit eine referentenbezogene Abstandsmessung verwendet (im Gegensatz zu der clausebezogenen bei Givón 1983a) – der Abstand zwischen zwei Erwähnungen desselben Referenten wird also über die Anzahl der dazwischenliegenden Referenten gemessen. Konkret wird die referentielle Distanz eines Referenten wie folgt berechnet:

- Bei Erstvorkommen eines Referenten wird der Wert auf den Maximalwert für den Referentenabstand gesetzt (hier: 60, s. 3.6.3, vgl. das Vorgehen bei Givón 1983a).
- Die Position des Referenten in der Referenten-Folge wird gespeichert (hier in einem Python-Dictionary mit Referenten-IDs als Schlüssel).
- Beim nächsten Vorkommen des Referenten wird die Distanz zwischen der aktuellen und der abgespeicherten Position berechnet und dem Referenten an der aktuellen Position der Wert zugeordnet.
- Der Eintrag der Position des Referenten im Dictionary wird auf die aktuelle Position gesetzt (Dictionary als Index letzter Positionen).
- Gruppen von Referenten werden dabei aufgesplittet, d. h. das Vorkommen eines Referenten in einer Gruppe wird wie ein Einzelvorkommen behandelt.

Diese Berechnung der referentiellen Distanz stellt sich in Tabelle 6.1 als Pseudocode folgendermaßen dar:

Pseudocode (referentielle Distanz)

```
Data: refs-dataset
Result: referential distance values
initialization;
max.distance = 60;
current position = 0;
while not at end of this text do
    read current ref;
    if last position for current ref exists then
        referential distance = min(current position - last position for current ref, max.distance);
    else
        referential distance = max.distance;
    end
    last position for current ref = current position;
    current position = current position + 1;
end
```

Tabelle 6.1: Pseudocode zur Berechnung referentieller Distanz

	DIST	REF	SEM	SYN	PRA	PHR	Text-ID
1	60	[1]	AG	S	FRAME	pronP	728
2	60	[5]	PAT	ATTR	FRAME	attrP	728
3	2	[1]	AG	S	TOP	zero	728
4	2	[5]	PAT	ATTR	REPEAT	attrP	728
5	2	[1]	AG	AGR	TOP	px	728
6	60	[2]	LOC	ADV	FOC	locNP	728

Report 6.3.2: Datensatz nach Feature-Construction (Referentielle Distanz)

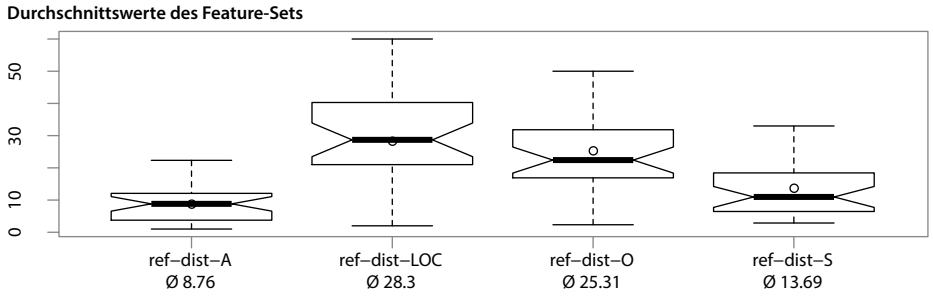
Wie bereits ausgeführt, wird hier also im Gegensatz zu der Operationalisierung der referentiellen Distanz bei Givón die Distanz eines Referenten zu seinem letzten Vorkommen nicht über deren Clause-Abstand berechnet, sondern – ermöglicht über die im Korpus vorliegende Referenten-Annotation – über den Referentenabstand, also darüber, wie viele Referenten zwischen dem letzten und dem aktuellen Vorkommen des Referenten liegen (inkl. Nullanapher, s. 3.6.3).

Der Maximalwert referentieller Distanz als Defaultwert für neue Referenten im Sinne des Schwellenwerts für Identifizierbarkeit wird hier als der dreifache bei Givón verwendete Wert (vgl. 1983a: 13f.) angesetzt (Annahme von durchschnittlich drei Referenten pro Clause, s. 3.6.3). Überschreitet die errechnete referentielle Distanz eines bereits bekannten Referenten diesen Maximalwert, liegt das letzte Vorkommen also weiter zurück, so wird der Distanzwert – analog zum Vorgehen bei Givón – abgeschnitten, also dieser Maximalwert gesetzt.

Feature-Extraction. Basierend auf dem im Feature-Construction-Prozess gebildeten Datensatz referentieller Distanzen, werden diese Daten nun aggregiert, d. h. in ein Bag-of-Tags-Feature-Set transformiert, indem als dessen Merkmale die Durchschnittswerte der referentiellen Distanz pro grammatischer Relation (S, A, O, LOC) in einem Text berechnet werden. Dabei wird zwischen Subjekt eines intransitiven Satzes (S = Subjective) und Subjekt eines transitiven Satzes (A = Agentive) über die im Datensatz abgespeicherte Information über den Transitivitätsstatus von Clauses differenziert.

Text-ID	ref-dist-A	ref-dist-LOC	ref-dist-O	ref-dist-S
728	12.08	39.30	31.17	7.00
730	3.29	40.29	2.33	13.59
732	2.68	23.69	22.83	7.59
741	21.33	60.00	22.00	3.00
742	3.67	50.33	31.88	22.03
750	9.08	33.75	28.65	7.08

Report 6.3.3: Feature-Set (Referentielle Distanz)



Plot 6.3.1: Boxplot des Feature-Sets (Referentielle Distanz)

Ergebnisse. Die untersuchten obugrischen Texte zeigen über alle Genres hinweg eine niedrigere referentielle Distanz im Subjektbereich (A und S, also transitives und intransitives Subjekt) als im Objektbereich, und vor allem als im adverbialen LOC-Bereich (s. Plot 6.3.1).

Die Unterschiede zwischen den Genres (s. Plot 6.3.2) sind am stärksten im adverbialen und im Objekt-Bereich: Die narrativen Texte weisen hier deutlich geringere Distanz-Werte auf als die übrigen Textsorten. Die insgesamt höheren Distanz-Werte für `jou` und `fas` (auch im Bereich intransitiver Subjekte) deuten darauf hin, dass in diesen informativen, nicht-narrativen Texten (ebenso wie bei den `eth`- und `myt`-Texten im lokativischen Bereich) – wie in 3.6.3 angenommen – weniger kontinuierliche Topiks auftreten, also schwächer topikale Referenten als in den narrativen Genres das Text-Modell konstituieren (insbesondere auch weniger *secondary topics* im O-Bereich, s. Nikolaeva 2001; vgl. 3.6.3). Stattdessen werden in diesen informativen Genres erwartbarerweise mehr neue Referenten eingeführt (vgl. 6.5.1), die bei ihrer Ersterwähnung jeweils den Maximalwert für referentielle Distanz erhalten. Insbesondere weisen die `jou`- und `fas`-Texte hohe referentielle Distanzwerte im S- und O-Bereich auf, also in den Argument-Positionen, die den bevorzugten Ort für Topik-Einführungen darstellen (s. 3.3.2; vgl. auch Du Bois 1987: 805: „Arguments comprising new information appear preferentially in the S or O roles, but not in the A role […]“).

Plot 6.3.2: GENRE-gruppierte Average-Scores-Barplots (Referentielle Distanz)

Die weitere Auswertung des Parameters referentieller Distanz erfolgt in 6.3.3 in einem kombinierten Feature-Set zusammen mit der Topik-Persistenz.

6.3.2 Topik-Persistenz

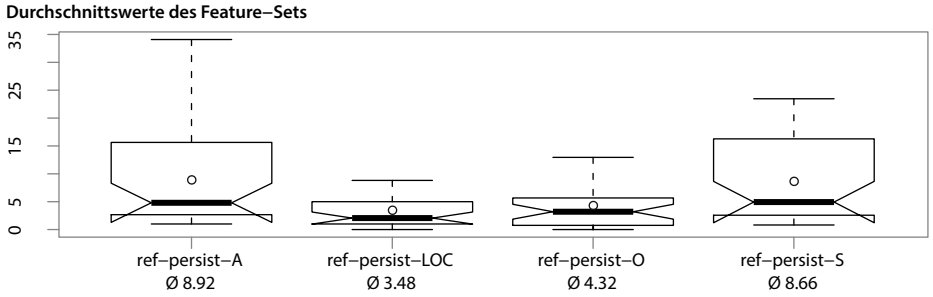
- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: grammatische Relationen
- **Feature-Wert**: Topik-Persistenz
- **Normierung der Features**: Textdurchschnitt pro Feature (GR)
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)
- **Ersatz NAs**: 0 (Minimalwert topikaler Persistenz)

Feature-Construction. Als die Anzahl der verbleibenden Erwähnungen eines Referenten relativ zu seiner Textposition wird die Topik-Persistenz wie folgt aus den basalen referentiellen Korpusdaten berechnet:

- Pro Text wird ein Index mit der verbleibenden Anzahl der Vorkommen der Referenten aufgebaut:
  - Dazu wird der Text in umgekehrter Sortierung referentenweise durchgegangen;
  - einem noch nicht in den Index aufgenommenen Referenten (= letzte Erwähnung im Text) wird der Wert 0 zugeordnet;
  - bei jedem weiteren Vorkommen eines Referenten wird der Wert im Index um 1 erhöht und dem aktuellen Referenten zugewiesen.
- Abschließend erfolgt eine umgekehrte Sortierung der Liste der getaggten Referenten.
- Im Korpus getaggte Gruppen von Referenten werden hier nicht aufgesplittet, sondern als eigenständige Einträge im mentalen Register behandelt (vgl. 3.6.3 und 3.6.4).

	PERSIST	REF	SEM	SYN	PRA	PHR	Text-ID
1	22	[1]	AG	S	FRAME	pronP	728
2	3	[5]	PAT	ATTR	FRAME	attrP	728
3	21	[1]	AG	S	TOP	zero	728
4	2	[5]	PAT	ATTR	REPEAT	attrP	728
5	20	[1]	AG	AGR	TOP	px	728
6	4	[2]	LOC	ADV	FOC	locNP	728

Report 6.3.4: Datensatz nach Feature-Construction (Topik-Persistenz)



Plot 6.3.3: Boxplot des Feature-Sets (Topik-Persistenz)

Feature-Extraction. Die im Feature-Construction-Prozess generierten Topik-Persistenz-Daten werden, analog zum Vorgehen für die referentielle Distanz, in ein Bag-of-Tags-Feature-Set mit Durchschnittswerten der Topik-Persistenz pro grammatischer Relation (S, A, O, LOC) und Textdokument transformiert.

Text-ID	ref-persist-A	ref-persist-LOC	ref-persist-O	ref-persist-S
728	7.00	1.30	0.67	6.50
730	4.43	1.82	1.00	8.00
732	15.63	5.19	8.61	16.27
741	1.33	0.00	1.00	2.00
742	5.00	0.89	0.75	4.59
750	22.08	11.62	14.35	23.08

Report 6.3.5: Feature-Set (Topik-Persistenz)

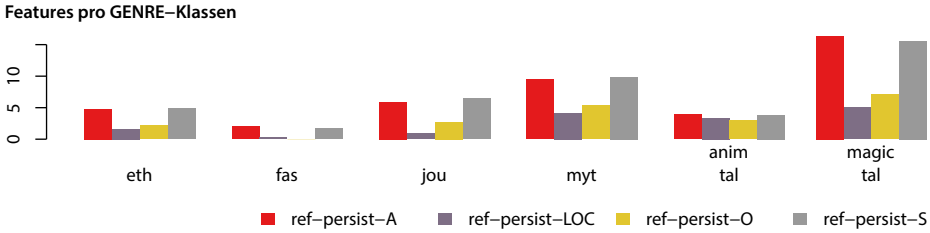
Ergebnisse. Die untersuchten obugrischen Texte zeigen für das Gesamtkorpus (s. Plot 6.3.3) eine höhere Topik-Persistenz im Subjektbereich (A und S) als im Objektbereich (O) und vor allem als im adverbialen Bereich (LOC); die Topik-Persistenz verhält sich damit erwartungsgemäß spiegelbildlich zu der referentiellen Distanz (vgl. Plot 6.3.4 mit Plot 6.3.2).

Insgesamt haben die Zaubermärchen (und ähnlich auch die mythologischen Texte) die mit Abstand höchsten Persistenz-Werte, insbesondere im S-A-Subjektbereich, was bei diesen längeren Texten aufgrund entsprechend umfangreicherer Topik-Kontinuitätsstrategien auch erwartbar ist.

Die kurzen Tiermärchen haben niedrigere Persistenz-Werte, insbesondere im S-A-Subjektbereich, in dem die Werte annährend gleich zum O- und LOC-Bereich liegen; dies lässt sich durch den für diese kurzen Tierfabeln typischen Text-Modell-Aufbau mit einem Handlungsort und wenigen Akteuren erklären, die aufgrund der sehr knappen Handlung nur kurz auftreten und entsprechend weniger persistent sind.

Die nicht-narrativen Texte haben relativ niedrige Persistenz-Werte, insbesondere im adverbialen Bereich (LOC), was sich mit den Feststellungen zur referentiellen Distanz oben spiegelbildlich deckt. Dies spricht dafür, dass in den narrativen Texten

eine stärkere Verortung der Referenten in der Textwelt im Sinne einer *cognitive map* mit wiederkehrenden Ortspunkten (*landmarks*) sowie auch durch weitere, im LOC-Bereich mit enthaltene adverbiale Angaben (Zeitangaben usw.) vorgenommen wird (vgl. Abschnitt 3.5).



Plot 6.3.4: GENRE-gruppierte Average-Scores-Barplots (Topik-Persistenz)

Die weitere Auswertung des Topik-Persistenz-Parameters erfolgt im nächsten Unterabschnitt als kombiniertes Feature-Set zusammen mit dem Parameter der referentiellen Distanz.

6.3.3 Kombiniertes Distanz-Persistenz-Modell

- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: grammatische Relationen
- **Feature-Werte**: referentielle Distanz, Topik-Persistenz
- **Normierung der Features**: s. Teilmodelle oben
- **Skalierung des Feature-Sets**: z-Standardisierung
- **Ersatz NAs**: s. Teilmodelle oben

Feature-Construction und -Extraction. Die Feature-Sets der referentiellen Distanz sowie der Topik-Persistenz werden hier in einem Feature-Set zusammengeführt, um auf diesem kombinierten Modell ana- und kataphorische Koreferenz gemeinsam zu untersuchen. Eine Feature-Construction entfällt dabei, es werden lediglich die bereits konstruierten Merkmale in einem Modell kombiniert. Im Gegensatz zu den Einzelmodellen ist hier allerdings eine Skalierung des Feature-Sets notwendig, da mit der Distanz-Skala (Abstand zum letzten Vorkommen des Referenten) und der Persistenz-Skala (Häufigkeit des Referenten im Resttext) zwei verschiedene Skalen auftreten. Als Skalierungsmethode wird die z-Standardisierung eingesetzt; man erhält folgen-

des Feature-Set (die skalierten Werte geben die Abweichung vom jeweiligen, auf 0 zentrierten Feature-Mittelwert an).[35]

Text-ID	ref-dist-A	ref-dist-LOC	ref-dist-O	ref-dist-S	ref-persist-A	ref-persist-LOC	ref-persist-O	ref-persist-S
728	0.56	0.68	0.41	−0.66	−0.20	−0.66	−0.78	−0.28
730	−0.92	0.74	−1.62	−0.01	−0.47	−0.50	−0.70	−0.09
732	−1.02	−0.29	−0.17	−0.60	0.70	0.51	0.91	0.99
741	2.10	1.96	−0.23	−1.06	−0.80	−1.05	−0.70	−0.87
742	−0.85	1.36	0.46	0.82	−0.41	−0.78	−0.76	−0.53
750	0.05	0.34	0.23	−0.65	1.38	2.45	2.13	1.87

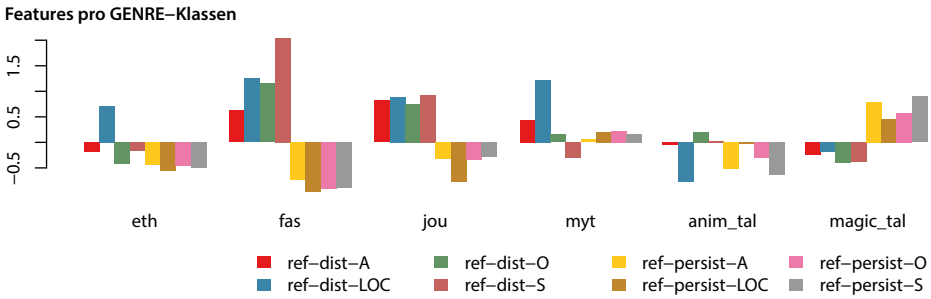
Report 6.3.6: Skaliertes Feature-Set (Distanz-Persistenz-Modell)

Ergebnisse. Eine genrebezogene Auswertung (s. Plot 6.3.5) des skalierten kombinierten Distanz-Persistenz-Feature-Sets zeigt noch einmal zusammenfassend, dass die kurzen Fabel-Tiermärchen im LOC-Bereich eine geringe referentielle Distanz aufweisen, also eine hohe durchschnittliche anaphorische Topikalität (vgl. auch die hohen Topikalitätsquotientenwerte für `anim_tal` in 6.3.5). Die kataphorische Topikalität (Topik-Persistenz) dieser Texte ist dagegen in allen Bereichen – aufgrund der Kürze der Texte – niedriger als bei den längeren Zaubermärchen, die sich über viele Abschnitte erstrecken und Referenten über eine längere Zeit wiederholt aufnehmen. Dieser Subtyp narrativer Texte im Korpus hat dafür nur mittlere Distanzwerte, also weniger anaphorisch-topikale Referenten. Im Persistenzbereich streuen diese narrativen Texte im LOC-Bereich relativ stark (s. Plot E.5): So haben z. B. die mythologische Sage 750 („The Middle Sosva Old Man's Raid to the Sacred Site on the Water", mit Bezug zu realen Orten: Kriegszug zum Fluß Lozva) sowie die beiden Mansi-Zaubermärchen 1262 und 1263 eine hohe LOC-Persistenz; auch hier spielen bestimmte Orte wie z. B. eine wiederholt aufgesuchte Hütte eine zentrale Rolle (vgl. Nguyen u. a. 2012: 380 bzgl. *domain knowledge* in Legenden; vgl. auch Cushing 1980: 228).

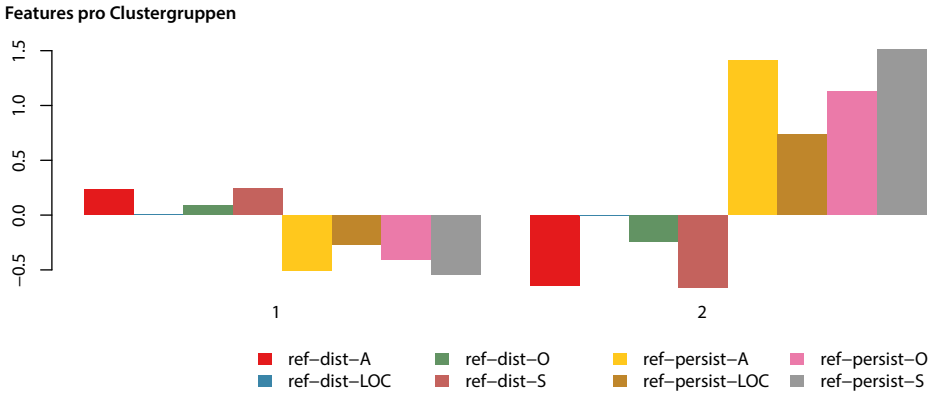
Die journalistischen Texte dagegen enthalten zwar eine Vielzahl an Ortsangaben (etwa eine Aufzählung besuchter Orte), diese werden aber nicht wieder regelmäßig aufgenommen, was sich in einer geringen Persistenz im LOC-Bereich ausdrückt. Allgemeiner ist festzustellen, dass die nicht-narrativen Texte im Korpus (`jou`, `fas` und eingeschränkt auch `eth`) gekennzeichnet sind durch eine relativ hohe referentielle

35 Wie oben bei den globalen Feature-Sets ist hier zu beachten, dass durch die Skalierung auf den Mittelwert und die Streuung jedes Features die ursprüngliche Skala verloren geht; ein Vergleich der Werte verschiedener Features (wie oben: `ref-persist-A` > `ref-persist-LOC`) ist hier also nicht mehr möglich. Die Unterschiede in den skalierten Features zeigen die Abweichung vom jeweiligen Mittelwert an und sind entsprechend zu interpretieren.

Distanz sowie eine niedrige Persistenz, also eine niedrige ana- sowie kataphorische Topikalität, insbesondere im LOC-Bereich.



Plot 6.3.5: GENRE-gruppierte Average-Scores-Barplots (Distanz-Persistenz-Modell)



Plot 6.3.6: Clustergruppierte Average-Scores-Barplots (Distanz-Persistenz-Modell)

Clustering (Plot E.1). Das Clustering zeigt eine binäre Abtrennung der längeren, ‚klassischen' Erzählungen (`magic_tal`) vom Rest; dazu gehören insbesondere die nordmansischen Zaubermärchen sowie der epische Mythos 750. Diese Clustergruppe 2 zeichnet sich aus (s. Plot 6.3.6) durch eine höhere durchschnittliche Topik-Persistenz (kataphorische Topikalität) sowie eine (leicht) niedrigere durchschnittliche referentielle Distanz (anaphorische Topikalität), insbesondere im S-A-Bereich – also durch eine insgesamt höhere phorische Topikalität der Referenten, die aus dem Aufbau eines Text-Modells mit einem mentalen Referenten-Register einer größeren Anzahl von regelmäßig wiederaufgenommenen Referenten resultiert (Topic-Continuity in längeren Texten). Ein Vergleich mit den einzelnen Feature-Sets referentieller Distanz und Topik-Persistenz zeigt, dass das Feature-Set der Topik-Persistenz hier den besten Cluster-Tendency-Wert erreicht (0.26 Hopkins).

Klassifikation. Das kombinierte Distanz-Persistenz-Feature-Set erreicht für die COMM_SIT-Kategorisierung einen Kappa-Wert > 0.6 (*substantial agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle. Die Feature-Importance-Werte (s. Plots E.6ff.) zeigen, dass für alle Textsortenklassen die Persistenz- und Distanzwerte von LOC-Referenten eine wichtige Rolle zur Klassendifferenzierung spielen. Für die COMM_SIT- und die DISC_STRUCT-Klassifizierung ist darüber hinaus die referentielle Distanz von O-Referenten relevant, für GENRE ist auch die Persistenz im Subjektbereich (A und S) ein weiteres wichtiges Merkmal zur Differenzierung, was mit der obigen Feststellung einer im Vergleich mit den längeren Zaubermärchen bedeutend niedrigeren Persistenz für A und S im narrativen Subgenre der Tiermärchen übereinstimmt, die in dieser Textsorteneinteilung als zwei getrennte Klassen auftreten.

6.3.4 Textweiter Topikalitätsquotient

Der Topikalitätsquotient als die relative Textfrequenz von Referenten wird im Folgenden für unterschiedliche Operationalisierungen von auf die Topikalitätsstärke bezogenen TWM-Parametern verwendet. Datengrundlagen für die Berechnung der Topikalitätsquotienten der einzelnen Referenten eines Textes sind wie bei den vorherigen Parametern die Daten der Referenten-Annotation (s. Report 6.3.1).

Feature-Construction. Allgemein berechnet sich der Topikalitätsquotient als die relative Häufigkeit der Vorkommen (Erwähnungen) eines Referenten im Text.

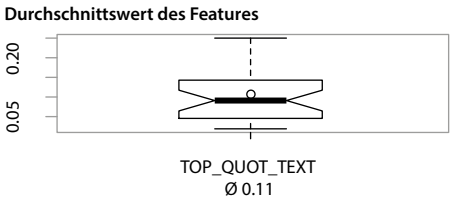
	Text-ID	REF	TOP_QUOT	REF_COUNT	ALL_REF_COUNT
1	728	[1]	0.36	23	64
2	728	[5]	0.06	4	64
3	728	[2]	0.08	5	64
4	728	[3]	0.08	5	64
5	728	[4]	0.02	1	64
6	728	[6]	0.03	2	64

Report 6.3.7: Datensatz nach Feature-Construction (Topikalitätsquotienten)

Ergebnisse. Eine auf diesen Daten der Feature-Construction aufbauende Kurzauswertung für den durchschnittlichen textweiten Topikalitätsquotienten, der der inversen Referentenanzahl (1/Referententypes) entspricht und damit als inverses Maß der absoluten referentiellen Informationsdichte verstanden werden kann (s. 3.6.2), zeigt, wie zu erwarten, für journalistische Texte (jou) die niedrigste durchschnittliche Topikalität, für die kurzen fabelartigen Tiermärchen (anim_tal) die höchste (s. Plot F.3).

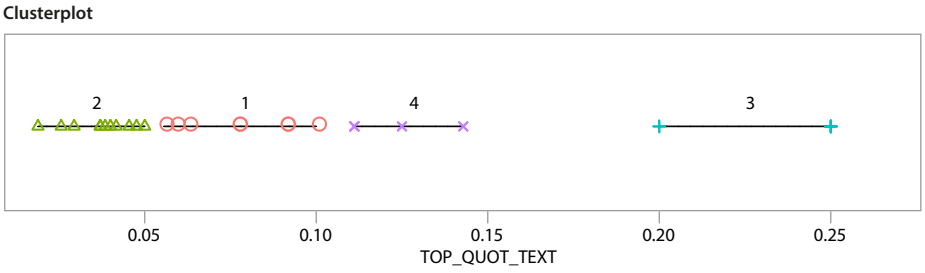
Text-ID	TOP_QUOT_TEXT
728	0.06
730	0.06
732	0.05
741	0.25
742	0.04
750	0.05

Report 6.3.8: Feature-Set (Textweiter Topikalitätsquotient)



Plot 6.3.7: Boxplot des Feature-Sets (Textweiter Topikalitätsquotient)

Clustering (Plot F.1). Ebenso zeigt sich im Clustering (s. Plot 6.3.8) mit dem textweiten Topikalitätsquotienten als Feature eine Trennung insbesondere der Kurzerzählungen (`anim_tal`; Gruppe 3) mit geringer Anzahl an Referententypes (höchster textweiter Topikalitätsquotient) von den übrigen Texten, wobei hier wiederum die Gruppe, die die journalistischen Texte enthält, die geringste durchschnittliche Topikalität hat (d. h. die höchste absolute Anzahl an Referenten); in dieser Gruppe 2 sind aber auch die längeren Zaubermärchen vertreten. Der textweite Topikalitätsquotient ist in dieser Operationalisierung als Maß der absoluten referentiellen Informationsdichte, das also ohne Längennormierung ein direktes Maß der ‚Größe' des konstruierten kognitiven Text-Modells (bzw. dessen Registers) ist, somit erwartbarerweise nicht geeignet, um allgemein narrative von nicht-narrativen Textsorten zu trennen. Dennoch kann dieser Parameter als kognitives Textlängenmaß des Umfangs der in einem Text-Modell beteiligten Referenten Relevanz für die gebrauchsbasierte Typisierung von Texten als Text-Weltmodell durch die Kognition haben (vgl. 6.6.1).



Plot 6.3.8: Clusterplot (Textweiter Topikalitätsquotient)

Klassifikation. Im Rahmen der Kurzauswertung des Merkmals textweiter Topikalitätsstärke wurde auf eine Klassifikation verzichtet.

6.3.5 Topikalitätsquotienten-Verteilung

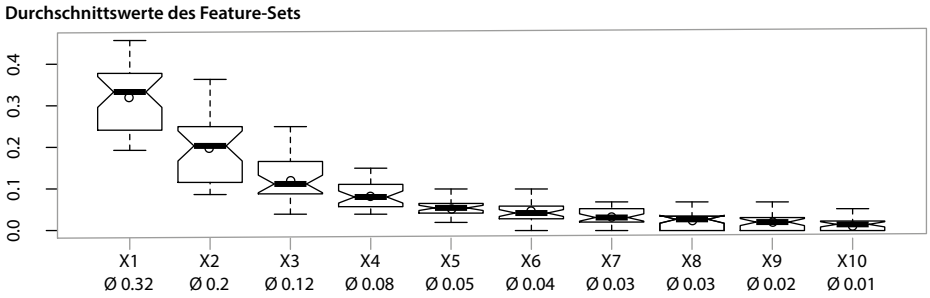
- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: nach Topikalitätsstärke sortierte Referenten
- **Feature-Wert**: relative Häufigkeit pro Referent (Topikalitätsstärke)
- **Normierung der Features**: relative Häufigkeit
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)

Feature-Construction. Die Konstruktion der Merkmale für das im Folgenden untersuchte Feature-Set relativer Textfrequenzen der häufigsten Referenten eines Textes basiert auf den in der Feature-Construction im vorherigen Abschnitt berechneten Daten zu den Topikalitätsquotienten der Referenten.

Feature-Extraction. Diese Daten werden zunächst so transformiert, dass ein Feature-Set mit den Referenten als Features und den Topikalitätsquotienten als Werten entsteht. Dieses vorläufige Feature-Set wird nun pro Text absteigend frequenzsortiert, sodass das erste Feature `X1` jeweils den häufigsten Referenten des Textes repräsentiert (den Referenten mit dem höchsten Topikalitätsquotienten), das zweite Feature `X2` den zweithäufigsten usw. Dieses Sortierung gewährleistet eine Vergleichbarkeit der Referenten-bezogenen Topikalitätswerte zwischen den Texten, die etwa bei Verwendung der im Rahmen der Annotation vergebenen Referenten-IDs als Merkmale nicht gegeben wäre. Als Feature-Set wird hier ein Subset der zehn topikalsten Referenten ausgewählt.

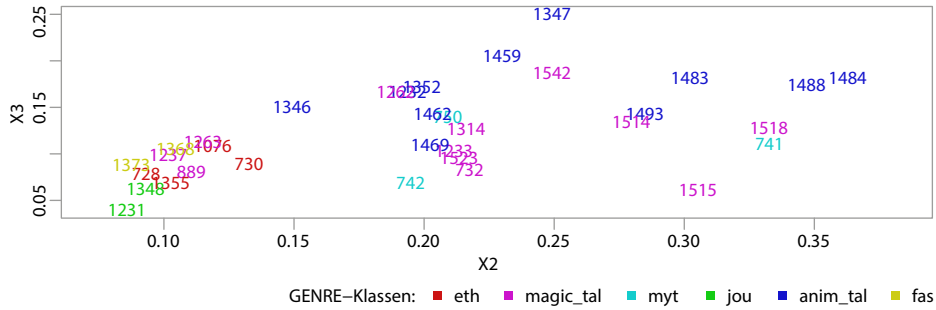
Text-ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
728	0.36	0.09	0.08	0.08	0.06	0.06	0.06	0.05	0.03	0.03
730	0.33	0.13	0.09	0.09	0.06	0.03	0.03	0.03	0.03	0.02
732	0.36	0.22	0.08	0.05	0.05	0.04	0.03	0.02	0.02	0.02
741	0.44	0.33	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00
742	0.24	0.20	0.07	0.06	0.06	0.06	0.06	0.03	0.03	0.02
750	0.24	0.21	0.14	0.11	0.04	0.04	0.03	0.03	0.03	0.02

Report 6.3.9: Feature-Set (Topikalitätsquotienten)



Plot 6.3.9: Boxplot des Feature-Sets (Topikalitätsquotienten)

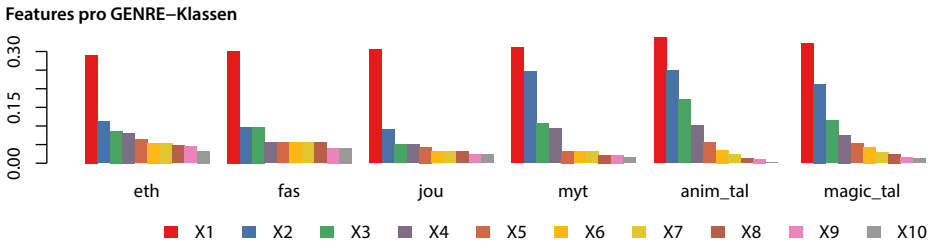
Ergebnisse. Die Durchschnittswerte der zehn topikalsten Referenten der untersuchten, schwerpunktmäßig narrativen obugrischen Texte zeigen eine kontinuierliche Abnahme der Topikalität um ca. 1/3, ausgehend von einem Hauptreferenten mit durchschnittlicher Topikalitätsstärke von 0.32 (s. Plot 6.3.9); im Durchschnitt beziehen sich also knapp 1/3 aller Referentenerwähnungen in einem Text auf den Hauptreferenten. Differenziert nach Textsorten (s. Plot 6.3.11) zeigt sich, dass die informativen Genres (`jou`, `fas`, `eth`) von dieser Topikalitätsverteilung insofern abweichen, als sie im Vergleich zum ersten Hauptreferenten (dem Thema des Textes; vgl. auch 3.6.1) für den zweitstärksten Referenten (`X2`) ebenso wie für die nachfolgenden Referenten deutlich geringere Topikalitätswerte aufweisen (vgl. Plot 6.3.10), dass dafür aber die Abnahme der Topikalität geringer ist, sich diese also gleichmäßiger auf die folgenden Referenten verteilt. Die narrativen Texte haben dementsprechend – im Gegensatz zu den informativen Texten – in der zweiten Hälfte des Feature-Sets (`X6-X10`) niedrigere Werte, die z. T. sogar den Wert 0 annehmen, wenn die Text-Modelle weniger als zehn Referenten enthalten.



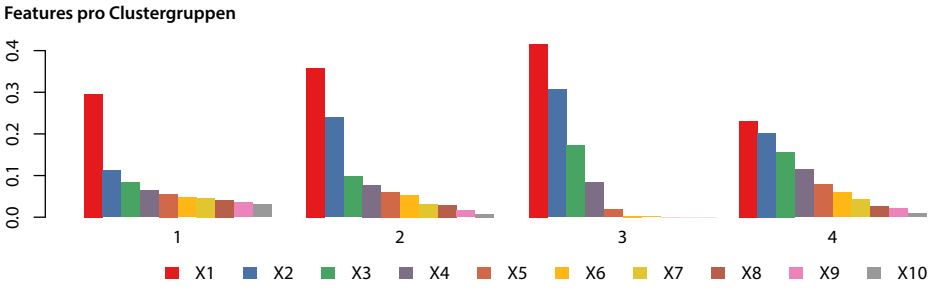
Plot 6.3.10: Topikalitätsstärke zweiter und dritter Referent nach GENRE-Klassen (Topikalitätsquotienten)

Clustering (Plot G.1). Es zeigt sich eine Clustertypologie mit jeweils disparater Topikalitätsverteilung (s. Plot 6.3.12):

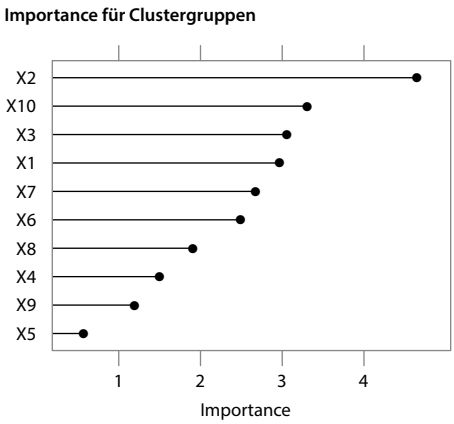
- Gruppe 1 enthält als Peripherie-Cluster (neben wenigen längeren Erzählungen) alle journalistischen, ethnographischen und lyrischen (also nicht-narrativen) Texte; es gibt einen relativ stark topikalen Referenten und eine größere Zahl mittel bis schwach topikaler Referenten.
- Gruppe 2 enthält verschiedene Erzähltexte; es gibt zwei relativ starke Hauptreferenten, danach fällt die Topikalität relativ gleichmäßig ab (zu dieser Gruppe gehört u. a. eine Variante des `TwoFires`-Märchens mit zwei Feuerstellen als Gegenspieler; ebenso die kürzere Surgut-Khanty-Variante 732 des `LittleBird`-Märchens mit den beiden Hauptprotagonisten des Vögleins und seiner Schwester, vgl. Csepregi 2005).
- Gruppe 3 enthält primär `anim_tal`-Kurzerzählungen; es gibt wenige stark topikale Referenten, die Topikalität fällt sehr steil ab.
- Gruppe 4 enthält überwiegend epische `magic_tal`-Erzählungen, aber auch einige Kurzerzählungen; es gibt hier keinen einzelnen, besonders topikalen Hauptreferenten, sondern eine größere Zahl relativ stark topikaler Referenten mit gleichmäßiger Abnahme der Topikalität vom ersten Referenten an.



Plot 6.3.11: GENRE-gruppierte Average-Scores-Barplots (Topikalitätsquotienten)



Plot 6.3.12: Clustergruppierte Average-Scores-Barplots (Topikalitätsquotienten)



Plot 6.3.13: Feature-Importance für Clustergruppen (Topikalitätsquotienten)

Das Clustering zeigt also einerseits Untergruppen innerhalb der Klasse narrativer Texte (Cluster 3–4), die sich durch einen unterschiedlichen Verlauf des Abfalls der Topikalität über die entsprechend sortierten Features unterscheiden (die kurzen Texte von Cluster 3 mit wenigen Referenten zeichnen sich insbesondere dadurch aus, dass die Topikalitätswerte ab der Mitte des Feature-Sets, also von `X6–X10`, gegen 0 gehen). Daneben wird mit Cluster 1 mit Texten primär nicht-narrativer, informativer Genres eine Gruppe abgetrennt, die sich u. a. durch niedrige `X2`-Topikalitätswerte auszeichnet, was auch der obigen Feststellung bei der gruppierten Auswertung nach Textsorten entspricht. Die Bedeutung der Topikalitätsstärke des zweithäufigsten Referenten (`X2`) für die Differenzierung der Clustergruppen zeigt sich auch in deren Feature-Importance-Analyse (Plot 6.3.13). Auch in der vergleichenden kombinierten Analyse des Gesamt-Feature-Sets in 6.6.1 stellt sich `X2` als wichtigstes (sowie `X10` als drittwichtigstes) differenzierendes Merkmal für die `BINARY`-Klassifizierung heraus, welche als binäre Textsorteneinteilung die narrativen von den nicht-narrativen Texten differenziert (s. Plot 6.6.9).

Klassifikation. Das Topikalitätsquotienten-Feature-Set erreicht für die `COMM_SIT`-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle. Die Feature-Importance-Werte (s. Plots G.7ff.) zeigen, dass der zweit-, dritt- und vierthäufigste Referent (Features `X2`, `X3`, `X4`) für alle Textsortenklassen eine wichtige Rolle zur Klassendifferenzierung spielt, z. T. auch der mit höchster (`X1`) sowie diejenigen mit niedrigster Frequenz (`X8–X10`).

6.4 Relationale Textstruktur-Typologie

	SEM_DOM	PHR	SYN	Text-ID
1	PERCEPT	compC	SUBPRED	728
2	MOTION	finVP	PRED	728
3	PERCEPT	compC	SUBPRED	728
4	MOTION	subC	SUBPRED	728
5	ACT	finVP	PRED	728
6	PERCEPT	finVP	PRED	728

Report 6.4.1: Relationaler Primärdatensatz (Relationale Parameter)

Der Aufbau relationaler Strukturmodelle als Operationalisierungen der **Ereignistypik** sowie der **Ereignisabfolge** kognitiver Text-Modelle basiert auf den semantischen Annotationsdaten zu Verbklassen (z. T. auch zu nominalen semantischen Domänen), die auf Grundlage der Daten des automatischen semantischen Taggings des Korpus generiert wurden (vgl. 5.3.3). Der relationale Primärdatensatz (Report 6.4.1) besteht aus den als prädikative Einheiten (`PRED` bzw. `SUBPRED`) annotierten Größen. Die folgenden fünf verbalen Hauptklassen werden dabei ausgezeichnet:

- `ACT` (`ACTION`): Handlungs- und Prozessverben
- `MOTION`: Bewegungsverben
- `PERCEPT` (`PERCEPTION`): Verben der Wahrnehmung
- `SPEECH`: Sprachhandlungsbezogene Verben
- `STATE`: Zustandsverben

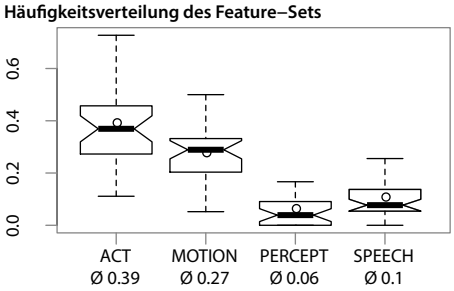
6.4.1 Ereignistypik

- **Beobachtungsgegenstand**: verbale Einheiten
- **Feature-Attribut**: Verbklassen
- **Feature-Wert**: Frequenzdaten der Verbklassen
- **Normierung der Features**: relative Textfrequenz
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)

Feature-Construction und -Extraction. Auf Grundlage des erstellten Ereignis-bezogenen Primärdatensatzes (Report 6.4.1) kann zunächst die Stärke der Verbklassen als Ereignistypen über deren relative Textfrequenz berechnet und aus diesen Daten ein Bag-of-Tags-Feature-Set erstellt werden (Report 6.4.2), wobei ein Subsetting auf die vier Aktivitäts-bezogenen verbalen Ereignistypen `ACT`, `MOTION`, `PERCEPT` und `SPEECH` vorgenommen wurde.

Text-ID	ACT	MOTION	PERCEPT	SPEECH
728	0.37	0.40	0.11	0.00
730	0.33	0.32	0.03	0.03
732	0.37	0.30	0.03	0.10
741	0.25	0.75	0.00	0.00
742	0.37	0.27	0.06	0.14
750	0.23	0.34	0.03	0.26

Report 6.4.2: Feature-Set (Ereignistypik)

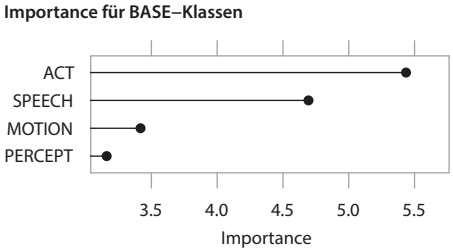


Plot 6.4.1: Boxplot des Feature-Sets (Ereignistypik)

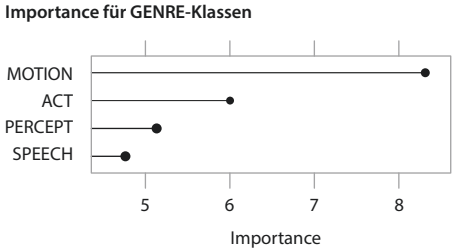
Ergebnisse. Die Ereignistypik der untersuchten, schwerpunktmäßig narrativen obugrischen Texte zeigt – in Übereinstimmung mit der Erwartung für narrative Texte als „stark motions- und handlungsbezogen" (Schulze 2020: 607) – eine **Dominanz des `ACTION`- und `MOTION`-Bereichs** (Plot 6.4.1), die zusammen 2/3 aller Ereignisvorstellungen ausmachen. In der gruppierten Analyse (s. Plot H.5) sieht man, dass sich im Gegensatz dazu die journalistischen Texte sowie die Fate Songs durch ein ausgewogenes Verhältnis der vier in diesem Feature-Set untersuchten Ereignistypen auszeichnen.

Clustering (Plot H.1). Im Clustering (s. Plot H.2) zeigt sich eine Hauptgruppe (1) mit zum Gesamtdurchschnitt ähnlicher Frequenzverteilung der Ereignistypen (jeweils ca. 1/3 `MOTION` sowie `ACTION`). Davon kann eine Gruppe (3) kurzer, stark `ACTION`-bezogener Texte unterschieden werden, die u. a. die `Cranberry`-Varianten enthält; der kurze Schöpfungsmythos 741 mit hohem `MOTION`-Anteil ist hier als Ausreißer anzusehen.

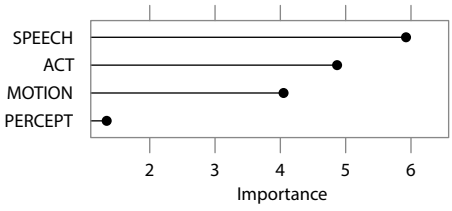
Klassifikation. Das Ereignistypik-Feature-Set erreicht für die `COMM_SIT`-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle. Die Feature-Importance-Werte (s. Plots 6.4.2ff.) zeigen, dass für die `GENRE`-Klassifizierung Handlungs- und Bewegungsverben die wichtigste Rolle für die Klassendifferenzierung spielen (vgl. die stark handlungsbezogene Gruppe 3 im Clustering mit primär kurzen narrativen `anim_tal`-Texten) und dass für die anderen Klassen jeweils (sprach-)handlungsbezogene Verben das wichtigste Merkmal bilden.



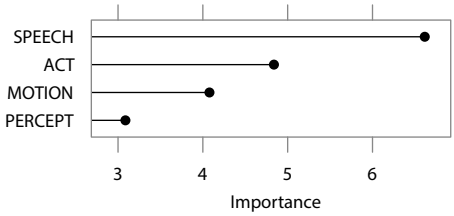
Plot 6.4.2: Feature-Importance für BASE-Klassen (Ereignistypik)



Plot 6.4.3: Feature-Importance für GENRE-Klassen (Ereignistypik)



Plot 6.4.4: Feature-Importance für COMM_SIT-Klassen (Ereignistypik)



Plot 6.4.5: Feature-Importance für DISC_STRUCT-Klassen (Ereignistypik)

6.4.2 Häufige Ereignisübergänge

- **Beobachtungsgegenstand**: verbale Einheiten
- **Feature-Attribut**: N-Gramme von häufigen Verbklassen-Übergängen
- **Feature-Wert**: Presence/Absence dieser Übergangs-N-Gramme
- **Normierung der Features**: keine (Presence-Absence-Zählung)
- **Skalierung des Feature-Sets**: keine (0/1-Kodierung)

Feature-Construction. Basierend auf dem relationalen Primärdatensatz (Report 6.4.1), können die häufigsten Abfolgemuster von Ereignistyp-Übergängen aus dem Korpus extrahiert werden (s. Report 6.1.2). Dazu wird zunächst für jeden Text die kategoriale Sequenzfolge der Verbklassen-Labels prädikativer Einheiten als Folgen von Ereignistyp-Zuständen mit der `seqdef`-Funktion des `TraMineR`-Pakets berechnet (Auflistung 6.29). Anschließend werden diese Ereignistyp-Zustandsfolgen mit der `seqecreate`-Funktion des `TraMineR`-Pakets in Sequenzen von Zustandsübergängen transformiert, die nur den Wechsel von Ereignistypen als Sequenzinformation behalten und damit von der Dauer eines Ereigniszustandes abstrahieren (Auflistung 6.30). Es entstehen neue, komplexe Sequenzlabels, die die Übergangsin-

formationen zwischen den Zuständen der Ausgangssequenz aufnehmen und damit selbst schon sequentielle Teilinformationen in sich tragen.

```
PERCEPT-MOTION-PERCEPT-MOTION-ACT-PERCEPT-MOTION-MOTION-MOTION-...
```

Auflistung 6.29: Zugrundeliegende Zustandsfolge (Häufige Ereignisübergänge)

```
(PERCEPT)-1-(PERCEPT>MOTION)-1-(MOTION>PERCEPT)-1-(PERCEPT>MOTION)-...
```

Auflistung 6.30: Zugrundeliegende Übergangsfolge (Häufige Ereignisübergänge)

Die Extraktion häufiger Ereignistyp-Übergänge aus diesem Datensatz von textweiten Übergangsfolgen erfolgt dann über die `seqefsub`-Funktion des `TraMineR`-Pakets (s. 6.1.2.6) mit folgenden Auswahlkriterien:

- `pmin.support`=0.67 (also Vorkommen in 2/3 aller Texte)
- `max.gap`=1 (zusammenhängende Teilsequenzen = N-Gramme, also keine Unterbrechung durch anderen Übergänge)
- `max.k`=3 (Maximallänge für die Übergangs-N-Gramme)

Die dabei durchgeführte Suche der häufigsten Übergangs-Teilsequenzen liefert als Resultat des Feature-Construction-Prozesses für häufige Ereignisübergange (s. Report 6.1.2) die fünf in mehr als zwei Drittel aller Texte des obugrischen Korpus vorkommenden Ereignisübergangsfolgen.

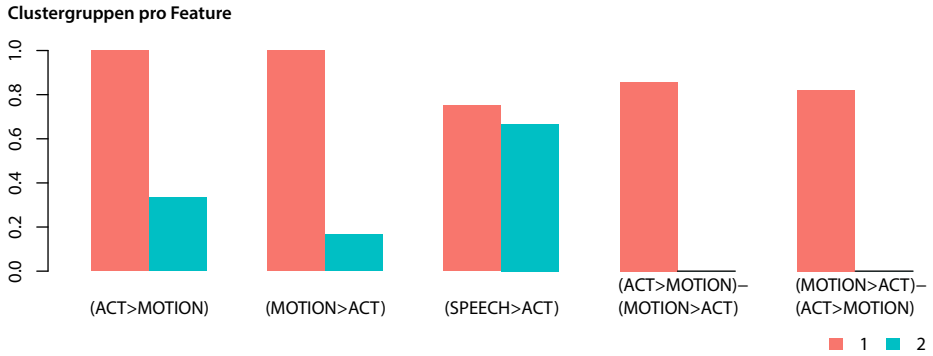
Feature-Extraction. Basierend auf diesen Frequent-Patterns als Merkmalen, wird nun ein Feature-Set (Report 6.4.3) erzeugt, indem mit der `seqeapplysub`-Funktion des `TraMineR`-Pakets für jeden Text das Vorkommen dieser Übergangsmuster (Presence-Absence-Zählung, vgl. 6.1.2.6) als Teilsequenzen in seiner globalen Sequenz von Ereignistyp-Übergängen überprüft wird.

Text-ID	(ACT>MOTION)	(MOTION>ACT)	(SPEECH>ACT)	(ACT>MOTION)-(MOTION>ACT)	(MOTION>ACT)-(ACT>MOTION)
728	1.00	1.00	0.00	1.00	1.00
730	1.00	1.00	1.00	1.00	1.00
732	1.00	1.00	1.00	1.00	1.00
741	1.00	1.00	0.00	0.00	1.00
742	1.00	1.00	1.00	1.00	0.00
750	1.00	1.00	1.00	1.00	1.00

Report 6.4.3: Feature-Set (Häufige Ereignisübergänge)

Ergebnisse. Die extrahierten häufigsten Ereignisabfolgen der untersuchten obugrischen Texte bestätigen die Annahme einer Dominanz von Übergängen zwischen und

unter Beteiligung von ACTION- und MOTION-Ereignissen; dazu gehört insbesondere auch die in Abschnitt 3.5 und in 3.7.3 als typische, mit *landmarks* assoziierte Abfolge in narrativen Texten festgestellte Ereignistyp-Übergangsfolge (MOTION>ACT)–(ACT>MOTION), also des Wechsels von einem Bewegungs- zu einem Handlungsblock und wieder zu einem Bewegungsblock (MOTION > ACTION > MOTION) in der textuellen Ereignisstruktur.



Plot 6.4.6: Featuregruppierte Average-Scores-Barplots der Clustergruppen (Häufige Ereignisübergänge)

Clustering (Plot I.1). Im Clustering wird eine kleine Gruppe (2) mit meist fehlenden ACTION <> MOTION-Übergängen und insbesondere ohne Ketten wie MOTION > ACTION > MOTION von einer Hauptgruppe (1) mit **starker Präsenz solcher ACTION <> MOTION-Übergänge** abgetrennt (s. Plot 6.4.6). Diese Kerngruppe kann, entsprechend der Erwartung an ein Volkserzählungs-TWM, als prototypischer Vertreter des obugrischen TWM-Parameters der Ereignisabfolge gelten. Wichtigste differenzierende Ereignisabfolge zwischen den beiden Clustergruppen ist nach der Random-Forest-Analyse (Plot I.3) der Übergang (MOTION>ACT).[36]

Klassifikation. Für das Frequent-Pattern-Feature-Set von häufigen Ereignisübergängen erreicht keine Kategorisierung einen Kappa-Wert > 0.4.

6.4.3 Ereignisabfolge-Sequenzen

- **Beobachtungsgegenstand**: verbale Einheiten
- **Sequenzdaten**: *tag*-Folge der Zustände bzw. Übergänge von Verbklassen
- **Normierung**: *maxlength/YujianBo* (bzw. *max*)

36 Während die Hopkins-Statistik für dieses Feature-Set, wohl aufgrund der Operationalisierung über Presence-Absence-Werte, mit 0.52 auf eine insgesamt schlechte Cluster-Tendency hinweist, deutet die Average-Silhouette-Width mit 0.56 bei zwei Clustern auf eine durchaus vorhandene Clustertrennung hin.

- **Distanzmaß**: OM/OM-like
- **Agglomerations-Methode**: Ward (Minimum Variance)

Die im Rahmen der Extraktion von häufigen lokalen Mustern von Ereignisübergängen erzeugten textweiten Zustands- und Übergangssequenzen (*transitions*) von Ereignistypen (im Folgenden mit `seq` bzw. `trans` abgekürzt) können auch direkt als globale sequentielle Repräsentation von Ereignisabfolgen im Rahmen einer kategorialen Sequenzanalyse Anwendung finden.

Sequenzextraktion. Dazu werden (wie bereits bei der Extraktion von Frequent-Tag-Patterns in 6.4.2) textweite kategoriale Sequenzfolgen (`seq`) der Verbklassen-*tags* aus den Annotationsdaten des relationalen Primärdatensatzes zur Ereignistypik (Report 6.4.1) generiert.

```
STATE-ACT-ACT-MOTION-ACT-...
```

Auflistung 6.31: Zustandsfolge: seq (Globale Ereignisabfolge)

Für die Transformation in Übergangssequenzen (`trans`) werden die Zustandsübergänge (wie oben in 6.4.2) aus den Zustandsfolgen extrahiert und textweite Folgen dieser Übergängen zwischen Ereignistyp-Zuständen erstellt.

```
(STATE)-1-(STATE>ACT)-2-(ACT>MOTION)-1-(MOTION>ACT)-...
```

Auflistung 6.32: Übergangsfolge: trans (Globale Ereignisabfolge)

Für die Klassifizierung der Sequenzen wird über die Berechnung von Optimal-Matching-Distanzen zwischen den globalen Sequenzrepräsentationen der Texte eine Distanzmatrix erstellt – für die Zustandssequenzen (`seq`) wird hier die `seqdist`-Funktion des `TraMineR`-Pakets verwendet, für die Folgen von Übergängen (`trans`; *transition events*) die `seqedist`-Funktion (mit OM-artigem Distanzmaß).[37]

Ergebnisse für Zustandsfolgen. Im nach Textsorten gruppierten Sequenz-Indexplot der globalen Zustandsfolgen (`seq`) von Ereignistypen der untersuchten obugrischen Texte (s. Plot 6.4.7) ist erkennbar, dass die narrativen (`anim_tal`, `magic_tal`) sowie die ethnographischen Texte (`eth`) von längeren `ACTION`-Blöcken geprägt sind, vor allem die Zaubermärchen (`magic_tal`) zusätzlich von einer Vielzahl von `MOTION`-Blöcken. Auffällig sind für die narrativen Genres auch Blöcke mit Wechseln von

37 Es werden die Default-Parameter von `seqedist` verwendet; für Details s. die Dokumentation unter https://www.rdocumentation.org/packages/TraMineRextras/versions/0.6.0/topics/seqedist (abgerufen am 31.10.2021). Bei Normierung mit `max` wird für die Übergangssequenzen ein ähnliches Clusteringresultat wie beim Clustering mit Zustandsfolgen erzielt.

SPEECH- und PERCEPT-Zuständen sowie ein STATE-geprägter Block als Exposition (Frame-Setting) zu Beginn des Textes.

Clustering von Zustandsfolgen (Plot J.1). Im Clustering (s. Plot J.3) der globalen Ereignis-Zustandsfolgen (seq) mit Ward-Agglomerationsmaß deckt sich die Differenzierung der Clustergruppen stark mit der Klasseneinteilung der Apriori-Kategorisierung der Kommunikationssituation (COMM_SIT):

- Gruppe 1 enthält überwiegend priv-Texte mit starkem ACTION-Anteil (insbesondere die Fabel-Tiermärchen und einige kürzere Yugan-Khanty-Zaubermärchen).
- Gruppe 2 enthält überwiegend publ-Texte mit ausgewogeneren Verhältnissen von MOTION, ACTION, PERCEPT und SPEECH (längere Nordmansi-Zaubermärchen, Fate Songs, journalistische Texte).

Entsprechend sind die durchschnittlichen Verweildauern der Clustergruppen (Plot J.2) den Verweildauern in den Zuständen in der COMM_SIT-Kategorisierung (Plot J.4) sehr ähnlich. Ein klares, die Texte einer Gruppe auszeichnendes Verlaufsprofil des Gesamtablaufs von Ereignistypen im Sinne von *event sequence patterns* (vgl. 3.7.2) ist hier schwer auszumachen; dies kann mit der Operationalisierung über alle PRED-Zustände (semantische Klassen von Prädikaten, also ggf. auch von Nomen) und einer entsprechend großen Anzahl an Zustandstypen zusammenhängen oder auch mit dem OM-Distanzmaß und der Längennormierung über *maxlength* (s. Diskussion dazu in 6.7.2).

Clustering von Übergangsfolgen (Plot J.5). Im Clustering von Ereignis-Übergangssequenzen (trans) mit Ward-Agglomerationsmaß ist festzustellen, dass die klassischen Zaubermärchen gemeinsam gruppiert werden (Cluster 2), abgetrennt von den Yugan-Khanty-Märchen, was darauf hindeutet, dass diese eine eigene Handlungsstruktur aufweisen. Auch werden, wie bei solchen Handlungsstruktur-bezogenen Merkmalen zu erwarten, Varianten wie die Cranberry-Märchen gemeinsam gruppiert, sowie auch die beiden Fate Songs, die als Beschreibungen von mit Orten, Dingen und Menschen verknüpften Empfindungen (vgl. Ojamaa & Ross 2004: 134) ebenfalls eine andersartige Handlungsstruktur aufweisen (textuell gekennzeichnet u. a. durch wiederholte Verwendung von Verben der Wahrnehmung).

Klassifikation. Zur Klassifikation der Ereignisabfolge-Sequenzen wird ein Spectrum-SVM-Klassifikator verwendet, der als Typ einer Feature-basierten Sequenzklassifikation eine Sequenz durch Teilsequenzen einer bestimmten Länge (Spektrum)

repräsentiert (s. 6.1.4.4).[38] Für die Spectrum-SVM-Klassifikation von Ereignis-Übergangsfolgen wird das mit dem `TraMineR`-Paket erzeugte Objekt der Übergangssequenzen in eine Liste von String-Sequenzen umgewandelt. Dazu werden die Level (Übergangstypen) der in R als Faktoren abgespeicherten Sequenzen als Alphabet verwendet.[39] Anschließend wird auf diesen *tag*-Folgen (1 Zeichen = 1 Zustand) mit 3-facher Kreuzvalidierung der Spectrum-SVM-Klassifikator trainiert und die Modell-Accuracy berechnet. Ähnlich wird für die Klassifikation der Zustandsfolgen vorgegangen.

In der Spectrum-SVM-Klassifikation erreichen beide sequentiellen Repräsentationsarten für globale Ereignisabfolgen (`trans` und `seq`) für die `COMM_SIT`-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle.

Plot 6.4.7: Sequenz-Indexplot für GENRE-Klassifizierung (Globale Abfolge Ereigniszustände)

38 Dies ähnelt dem Vorgehen bei der Modellierung von Ereignisabfolgen durch Frequent-Tag-Patterns (s. 6.4.2), insofern in beiden Fällen eine Menge von Teilsequenzen zur Klassifikation verwendet wird.

39 Bei der Spectrum-SVM-Klassifikation mit Übergängen als Sequenzzuständen ist ein Ersatz mit ASCII-Buchstaben-Alphabet, der für die Spectrum-SVM-Klassifikation einfacher *tag*-Zustandsfolgen mit begrenztem Vokabular an Labels wie z. B. in 6.5.6 Anwendung finden kann (vgl. 6.1.4.4), aufgrund der Menge an möglichen Übergangstypen problematisch.

6.5 Pragmatisch-informationsstrukturelle Modelle

Als referentenbezogene informationsstrukturelle TWM-Parameter werden u. a. eine regionale Frequenzanalyse der **Einführung neuer Topiks** als Operationalisierung von relativer Informationsdichte und -fluss sowie eine Sequenzanalyse von **Switch-Reference**-Verläufen als Operationalisierung textueller Perspektivierungsstrukturierung ausgewertet. Zu den untersuchten informationsstrukturbezogenen TWM-Parametern auf relationaler Ebene gehört die Background-Foreground-Strukturierung in einer Modellierung über **Subordinationsstrukturen** als Backgrounding-Strategie sowie über **Temporal-Sequencing** als Foregrounding-Strategie ebenso wie das Diskurslayering in Texten über eingebettete **Dialoge** als mesostruktureller Parameter der Informationsstrukturierung von Text-Modellen.

6.5.1 Topik-Einführungen

- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: regionale Stärke Topik-Einführung
- **Feature-Wert**: Frequenzdaten der Topik-Einführungen
- **Normierung der Features**: relative Regionen-Frequenz
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)

Feature-Construction. Grundlage der Feature-Construction für die regionale Stärke von Topik-Einführungen ist das parallel zur Berechnung der referentiellen Distanz (s. 6.3.1) auf dem referentiellen Primärdatensatz (Report 6.3.1) berechnete `INFO`-Merkmal (s. Report 6.5.1) mit den folgenden hier relevanten Werten:

- `new`: bei noch nicht im Index vorhandenen Referenten (Ersterwähnung)
- `given`: in allen übrigen Fällen[40]

	INFO	REF	SEM	SYN	PRA	PHR	Text-ID
1	new	[1]	AG	S	FRAME	pronP	728
2	new	[5]	PAT	ATTR	FRAME	attrP	728
3	given	[1]	AG	S	TOP	zero	728
4	given	[5]	PAT	ATTR	REPEAT	attrP	728
5	given	[1]	AG	AGR	TOP	px	728
6	new	[2]	LOC	ADV	FOC	locNP	728

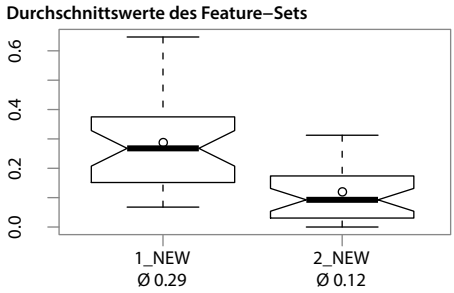
Report 6.5.1: Referentieller Datensatz nach Feature-Construction (Topik-Einführungen)

40 Im Rahmen der Feature-Construction wurde zunächst – analog zur Berechnung der referentiellen Distanz – bei Erreichen des Schwellenwerts für Identifizierbarkeit (Distanz von 60 Referenten) ein zusätzliches Tag (`before`) für wiederaufgenommene Referenten vergeben. In der endgültigen Auswertung des Anteils neuer Topiks wird diese Subdifferenzierung gegebener Referenten aber nicht berücksichtigt, neue Topiks werden strikt im Sinne einer Text-Givenness berechnet, vgl. 3.8.1.

Feature-Extraction. Basierend auf diesen im Feature-Construction-Prozess generierten Daten zu Topik-Einführungen, wird ein nach zwei Regionen differenziertes Feature-Set, mit den jeweiligen Textfrequenzen der Topik-Einführungen als Merkmale, als einfache Modellierung des Informationsflusses im Sinne einer Änderung der Informationsdichte zwischen erster und zweiter Texthälfte erstellt (Report 6.5.2). Dazu wird eine Textpartitionierung vorgenommen und die relative Häufigkeit neuer Topiks *pro Textregion* als Feature bestimmt (regionale Bag-of-Tags).

Text-ID	1_NEW	2_NEW
728	0.22	0.25
730	0.29	0.07
732	0.15	0.13
741	0.75	0.20
742	0.35	0.18
750	0.10	0.02

Report 6.5.2: Feature-Set für 2 Regionen (Topik-Einführungen)



Plot 6.5.1: Boxplot des Feature-Sets (Topik-Einführungen)

Ergebnisse. Das Feature-Set der auf zwei Regionen differenzierten Stärke von Topik-Einführungen der untersuchten obugrischen Texte zeigt – wie für die überwiegend narrativen Texte zu erwarten (s. Schulze 2020: 607; vgl. Abschnitt 3.8) – eine mehr als doppelt so hohe Anzahl an Topik-Einführungen in der ersten Texthälfte (`1_NEW`) gegenüber der zweiten (`2_NEW`). Konkret werden im Durchschnitt in der erste Texthälfte durch knapp 30% der Referentenerwähnungen neue Topiks eingeführt, in der zweiten Hälfte nur noch durch 12% (s. Plot 6.5.1).

In Übereinstimmung mit der Annahme einer prototypisch niedrigen Informationsdichte in narrativen Texten sowie einer relativ starken Abnahme des Informationsflusses bei einer Aktantenstruktur mit wenigen Hauptreferenten (s. Abschnitt 3.8) zeigt sich in der gruppierten Auswertung nach Genres (Plot K.6), dass für die narrativen Texte – insbesondere für die längeren Erzählungen des Zaubermärchen-Subgenres (`magic_tal` mit 18% in der ersten Texthälfte und 7% in der zweiten) – eine sehr niedrige Informationsdichte gegeben ist (s. auch Report 6.5.3 und 6.5.5). Dagegen haben die informativen, nicht-narrativen Texte (`fas`, `jou`, `eth` und auch `myt`) in beiden Texthälften bedeutend höhere `NEW`-Werte, es liegt bei diesen Texten also eine höhere relative referentielle **Informationsdichte** vor; vgl. hierzu auch die von Biber für die ersten 200 Wörter der Texte eines Korpus festgestellten Werte (1992b: 218, 232) eines Anteils neuer Referenten in informativen Texten von 65%, in Konversationen von 29%.

BASE	1_NEW	2_NEW
eth	0.33	0.22
fas	0.71	0.41
jou	0.42	0.23
myt	0.40	0.13
tal	0.22	0.07

Report 6.5.3: Durchschnittswerte in BASE-Klassen (Topik-Einführungen)

BASE	1_NEW	2_NEW
eth	0.10	0.12
fas	0.07	0.01
jou	0.04	0.09
myt	0.27	0.08
tal	0.10	0.06

Report 6.5.4: Standardabweichung in BASE-Klassen (Topik-Einführungen)

GENRE	1_NEW	2_NEW
eth	0.33	0.22
fas	0.71	0.41
jou	0.42	0.23
myt	0.40	0.13
anim_tal	0.26	0.06
magic_tal	0.18	0.07

Report 6.5.5: Durchschnittswerte in GENRE-Klassen (Topik-Einführungen)

GENRE	1_NEW	2_NEW
eth	0.10	0.12
fas	0.07	0.01
jou	0.04	0.09
myt	0.27	0.08
anim_tal	0.09	0.07
magic_tal	0.09	0.04

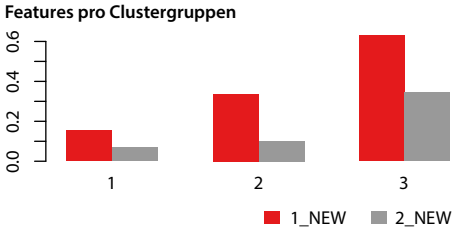
Report 6.5.6: Standardabweichung in GENRE-Klassen (Topik-Einführungen)

Auch die Abnahme der Werte zwischen den Texthälften ist in diesen nicht-narrativen Genres schwächer (nur 33% Abnahme für `eth`, 42% für `fas` und 45% für `jou` gegenüber 61% für `magic_tal` und 77% für `anim_tal`). Für den **Informationsfluss** in diesen informativen Texten gilt also, dass die relative referentielle Dichte über den Verlauf des Textes hin konstanter bleibt als in narrativen Texten (vgl. die Genre-Durchschnittswerte in den Plots K.5ff.; s. auch Plot K.4). Damit zeigt sich, dass in den narrativen Texten des obugrischen Korpus sowohl eine **geringere Informationsdichte** als in den nicht-narrativen, informativen Texten vorherrscht, als auch ein stärker **abnehmender Informationsfluss** (vgl. auch `2_NEW` als zweitwichtigstes differenzierendes Merkmal neben `X2` in der binären Apriori-Unterscheidung von narrativen und nicht-narrativen Texten `BINARY` in 6.6.1).

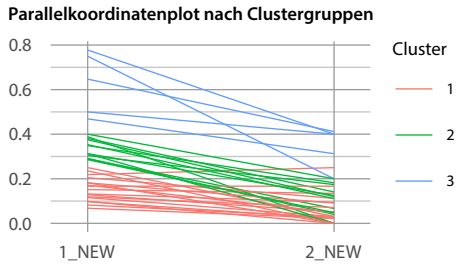
Clustering (Plot K.1). Im Clustering zeigen sich drei klar getrennte Gruppen (s. Plots K.2f.):

- Eine kleine Gruppe (3) peripherer, nicht-narrativer Genres (beide `fas`-Texte, je ein `jou`-, `eth`- und `myt`-Text) mit sehr hohen New-Topic-Werten und im Durchschnitt moderat fallender Tendenz in der zweiten Hälfte (s. Plots 6.5.2f.).
- Gruppe 2 mit dem Gesamtdurchschnitt ähnlichen Werten und im Durchschnitt stärkerer Abnahme von Topik-Einführungen in der zweiten Texthälfte; heterogener Genre-Mix (zur Hälfte kurze Fabelmärchen sowie die kürzeren Yugan-Khanty-Zaubermärchen, aber u. a. auch ein `jou`-Text).

- Gruppe 1 als großer Kerncluster narrativer Texte, vorwiegend aber die längeren `magic_tal`-Zaubermärchen mit niedrigeren und im Vergleich mit Gruppe 2 im Durchschnitt schwächer abnehmenden Topik-Einführungswerten.



Plot 6.5.2: Clustergruppierte Average-Scores-Barplots (Topik-Einführungen)



Plot 6.5.3: Parallelkoordinatenplot nach Clustergruppen (Topik-Einführungen)

Die Abtrennung der Clustergruppe 3 mit ausschließlich nicht-narrativen Texten mit den höchsten Topik-Einführungswerten deutet darauf hin, dass diese regionale Operationalisierung von relativer referentieller Informationsdichte und -fluss einen geeigneten TWM-Parameter für die Differenzierung der Informationsstruktur von Volkserzählungen gegenüber anderen Genres darstellen kann.

Klassifikation. Das regionale Topik-Einführungs-Feature-Set erreicht für die `COMM_SIT`- und die `DISC_STRUCT`-Kategorisierung mit zwei Regionen einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle. Für keine Textsortenklassifizierungen ist ein bedeutender Unterschied in der Feature-Importance zwischen erster und zweiter Texthälfte festzustellen.

6.5.2 Switch-Reference-Sequenzen

- **Beobachtungsgegenstand**: nominale Einheiten
- **Sequenzdaten**: *tag*-Folge von Switch-Reference-Werten
- **Normierung**: *maxlength*
- **Distanzmaß**: OM, euklidisch
- **Agglomerations-Methode**: Ward (Minimum Variance)
- **Klassifikationsmethode**: SVM, knn

Basierend auf den Annotationssdaten zu Referenten-IDs und syntaktischer Funktion im referentiellen Primärdatensatz (Report 6.3.1), wird die Switch-Subject-Struktur durch eine globale kategoriale Folge entsprechender Zustandslabels operationalisiert, um ein textweites Verlaufsprofil des Referentenwechsels in Subjektposition als sequentielles Modell der textuellen Perspektivierungsstruktur zu erhalten.

Feature-Construction. Folgende Werte werden als Switch-Reference-Status für alle Subjekte eines Textes durch Abgleich der Referenten-ID des aktuellen mit der des vorherigen Subjekts berechnet (Report 6.5.7):

- `CONT` bei gleichbleibendem Referenten in Subjektposition
- `SWITCH` bei Wechsel des Referenten in Subjektposition (sowie für das erste Subjekt im Text)
- `SPEECH` bei direkter Rede (als Unterbrechung der narrativen Handlungsebene und damit Aussetzung von Topic-Continuity-Strategien)

	SWITCH_STATUS	REF	SEM	CONSTR	TRANS	INFOSPEECH	Text-ID
1	SWITCH	[1]	AG	finVP	transitiv	None	728
2	CONT	[1]	AG	finVP	transitiv	None	728
3	CONT	[1]	AG	finVP		None	728
4	CONT	[1]	AG	finVP		None	728
5	CONT	[1]	AG	finVP		None	728
6	SWITCH	[3]	AG	finVP		None	728

Report 6.5.7: Referentieller Datensatz nach Feature-Construction (Switch-Reference-Sequenzen)

Sequenzextraktion. Als kategoriale sequentielle Repräsentation der Switch-Reference-Struktur von Texten werden diese `SWITCH_STATUS`-Werte mit der `seqdef`-Funktion des `TraMineR`-Pakets in entsprechende globale (textweite) *tag*-Folgen transformiert (s. Auflistung 6.33).

```
SWITCH-SWITCH-SWITCH-SWITCH-SWITCH-CONT-SPEECH-SPEECH-...
```

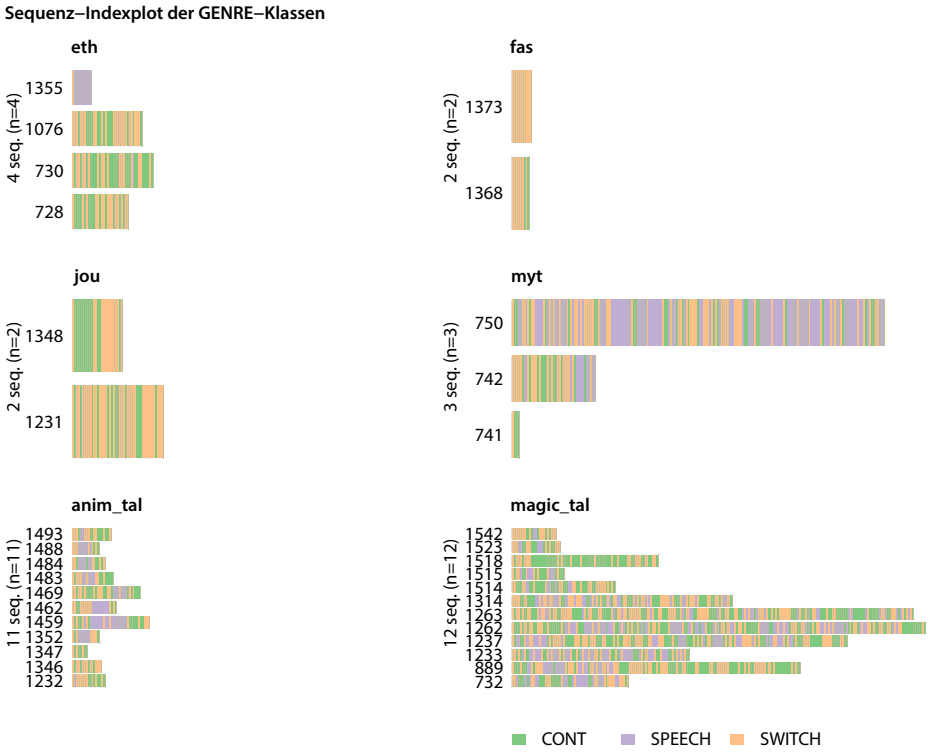
Auflistung 6.33: Zustandsfolge (Switch-Reference-Sequenzen)

Für die Klassifizierung dieser Folgen wird anschließend über die `seqdist`-Funktion des `TraMineR`-Pakets eine Optimal-Matching-Distanzmatrix erstellt, indem paarweise Abstände zwischen den Sequenzrepräsentationen der Textdokumente berechnet werden.

Ergebnisse. Eine deskriptiv-statistische Beurteilung der nach Genres geordneten Switch-Reference-Sequenzen im Indexplot (s. Plot 6.5.4) zeigt folgendes Ergebnis: Die Texte der peripheren Genres (`jou` und `fas`) sind geprägt durch wiederholten Wechsel des Referenten in Subjektposition (`SWITCH`) ohne Unterbrechung der primären Textebene durch Phasen direkter Rede (`SPEECH`); es gibt nur kurze Phasen von Topic-Continuity (`CONT`).
Die Perspektivierungsstruktur der narrativen Texte entspricht der erwartbaren Prototypik für Volkserzählungen einer **blockweisen Topic-Continuity**: Längere kontinuierliche Phasen (`CONT`-Textabschnitte mit demselben Referenten in Subjektposition), gefolgt von Switch-Reference als Perspektivwechsel (Wechsel des Diskurstopiks in

Subjektposition), unterbrochen von Dialog-Blöcken (s. 3.8.2; vgl. auch Schulze 2019: 29).

Die ethnographischen Texte weisen (mit der Ausnahme von Text 1355, `Trap`, s. dazu auch 6.7.1) eine ähnliche Perspektivierungsstrukturierung wie die narrativen Texte auf, allerdings ohne Dialogphasen (vgl. auch 6.5.6).



Plot 6.5.4: Sequenz-Indexplot für GENRE-Klassifizierung (Switch-Reference-Sequenzen)

Clustering (Plot L.1). Im Clustering mit Ward-Agglomerationsmaß zeigen sich folgende Gruppen (s. Plot L.2):

- Gruppe 2 enthält die längeren Mansi-Zaubermärchen mit vielen Dialogen und längeren kontinuierlichen Phasen.
- Gruppe 3 enthält eine Untergruppe der kurzen `anim_tal`-Fabel-Tiermärchen mit Dialog-Mittelteil.
- Gruppe 1 enthält als große Restgruppe narrative und nicht-narrative Genres ohne (oder mit sehr wenig) Dialogen.

Klassifikation. Für die Spectrum-SVM-Klassifikation werden die verschiedenen Zustände des Switch-Reference-Status (`SWITCH`, `CONT`, `SPEECH`) in Zeichen der

Länge 1 recodiert (s. 6.1.4.4) und diese als Alphabet verwendet. Anschließend wird auf diesen *tag*-Folgen mit 3-facher Kreuzvalidierung der Spectrum-SVM-Klassifikator trainiert. In der Spectrum-SVM-Klassifikation erreichen die Switch-Reference-Sequenzdaten allerdings in keiner Kategorisierung die Schwelle von Kappa > 0.4 (*moderate agreement*; der beste Wert wird für `GENRE` mit 0.34 erreicht).

6.5.3 Fokussierungstypik

- **Beobachtungsgegenstand**: nominale Einheiten
- **Feature-Attribut**: pragmatische Kategorien
- **Feature-Wert**: Frequenzdaten der pragmatischen Kategorien
- **Normierung der Features**: relative Textfrequenz
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)

Feature-Construction und -Extraction. Basierend auf den pragmatischen Annotationsdaten von Referenten im referentiellen Primärdatensatz, wird die textweite Stärke dieser pragmatischen referentiellen Kategorien als deren relative Textfrequenz berechnet (s. Report 6.5.8) und diese Daten in ein Bag-of-Tags-Feature-Set transformiert (s. Report 6.5.9).

	INFO	REF	SEM	SYN	PRA	PHR	Text-ID
1	new	[1]	AG	S	FRAME	pronP	728
2	new	[5]	PAT	ATTR	FRAME	attrP	728
3	given	[1]	AG	S	TOP	zero	728
4	given	[5]	PAT	ATTR	REPEAT	attrP	728
5	given	[1]	AG	AGR	TOP	px	728
6	new	[2]	LOC	ADV	FOC	locNP	728

Report 6.5.8: Referentieller Datensatz nach Feature-Construction (Fokussierungstypik)

Text-ID	FOC	CTR	REPEAT	MFOC	FRAME	TOP	THL
728	0.33	0.08	0.12	0.00	0.03	0.39	0.00
730	0.31	0.08	0.17	0.00	0.03	0.39	0.00
732	0.23	0.14	0.20	0.03	0.02	0.35	0.00
741	0.11	0.00	0.11	0.00	0.33	0.44	0.00
742	0.39	0.01	0.10	0.00	0.01	0.39	0.00
750	0.17	0.19	0.05	0.02	0.03	0.46	0.00

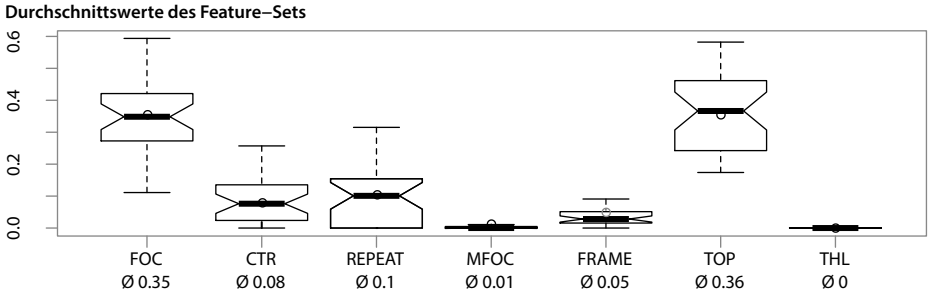
Report 6.5.9: Feature-Set (Fokussierungstypik)

Ergebnisse. Die pragmatische Typik (Fokussierungstypik für Aufmerksamkeitssteuerung) der untersuchten obugrischen Texte zeigt 5% `FRAME`-Setting und ca. 20% Fokusmarkierung (pronominal durch `CTR` oder durch nominale Wiederholung

= REPEAT). Zero-Formen (TOP) sowie nominale Ersterwähnungen (FOC) machen jeweils ca. 1/3 der Referentenerwähnungen aus (s. Plot 6.5.5).

Textsortenspezifisch zeigt sich (s. Plots M.4f.) ein relativ hoher FRAME-Anteil bei fas- und myt-Texten, also den expositorischen Diskursen (expos) mit logischer Abfolge eines Themas (vgl. 5.2.4).[41] Die prozeduralen Diskurse proc (mit chronologischer Themenfolge; z. B. Text 1076: MakeBread) haben einen hohen Anteil an REPEAT, also Fokussierung über nominal realisierte Wiederholung eines eingeführten Referenten (Fokusmarkierung durch unmittelbare Wiederholung, s. 5.3.5). Kontrastiver Fokus (CTR; identifiziert über Pronomen, s. 5.3.5) ist bei den expositorischen Texten schwach vertreten, sonst relativ gleichmäßig.

Die narrativen Texte zeichnen sich aus durch kurzes Frame-Setting sowie regelmäßige Fokussierung von Referenten (CTR und REPEAT gleichauf, jeweils 10%). Dies korrespondiert mit der erwarteten Prototypik für Volkserzählungen, nämlich der „Emphase markierter Situationen und ihrer Akteure" (Schulze 2020: 607).



Plot 6.5.5: Boxplot des Feature-Sets (Fokussierungstypik)

Clustering (Plot M.1). Im Clustering zeigen sich (neben dem kurzen Schöpfungsmythos 741 als Ausreißer) zwei Gruppen mit nur leicht unterschiedlicher pragmatischer Typik (s. Plot M.3):

- Gruppe 2 enthält u. a. die jou- und fas-Texte und vor allem anim_tal-Texte; sie hat einen durchschnittlich leicht höheren Fokussierungsgrad.
- Gruppe 1 enthält primär die längeren Zaubermärchen (magic_tal); sie hat einen höheren Satz-Topik-Anteil (TOP, im OUDB-Korpus als pragmatisches Label über Nullanaphern bestimmt).

Klassifikation. Das Fokussierungstypik-Feature-Set erreicht mit 0.59 für die COMM_SIT-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Die restlichen Kategorisierungen bleiben unter dieser Schwelle.

41 Vgl. z. B. markiertes Frame-Setting über Diskurspartikel im mythologischen Text Fireflood (742), Satz 1: „holy fireflood EMPH1 (*ta*) occurred".

6.5.4 Temporal-Sequencing

- **Beobachtungsgegenstand**: verbale Einheiten
- **Feature-Attribut**: Temporal-Sequencing-Stärke pro Text
- **Feature-Wert**: Frequenzdaten zu Temporal-Sequencing
- **Normierung der Features**: relative Häufigkeit
- **Skalierung des Feature-Sets**: keine (nur ein Feature)

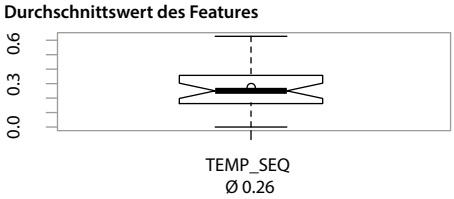
Das sog. Temporal-Sequencing als TAM-Foregrounding narrativer Sätze (s. 3.8.4) ist ein auf die relationale Informationsstrukturierung bezogener Parameter. Für das Obugrische wird hier Präsens-basiertes TAM-Foregrounding untersucht, also dessen Gebrauch als historisches Präsens, vgl. Nikolaeva 1999: 26 für das Khanty: „The Non-Past […] can also be used as the narrative past, especially in folklore texts."

Feature-Construction. Basierend auf dem relationalen Primärdatensatz wird der Parameter der Temporal-Sequencing-Stärke berechnet (s. Report 6.5.10), indem ein binäres `TEMP_SEQ`-Feature als Merkmal narrativer Sätze aus den Daten zum Tempus der Verbalformen sowie der Information bzgl. des direkte-Rede-Status konstruiert wird. Das Merkmal gilt als erfüllt (Wert = 1), wenn die Verbform im Präsens steht und nicht Teil direkter Rede ist (also auf der primären Handlungsebene).

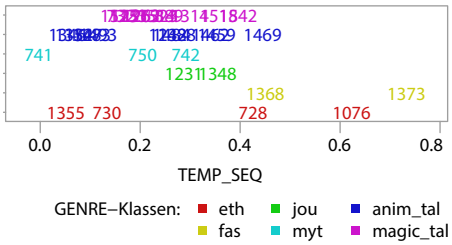
	TEMP_SEQ	INFOSPEECH	SEM_DOM	PHR	SYN	TEMP_PRED	Text-ID
1	0	None	PERCEPTION	compC	SUBPRED		728
2	0	None	MOTION	finVP	PRED		728
3	0	None	PERCEPTION	compC	SUBPRED		728
4	0	None	MOTION	subC	SUBPRED		728
5	0	None	ACTION&PROCESS	finVP	PRED		728
6	1	None	PERCEPTION	finVP	PRED	PRS	728

Report 6.5.10: Relationaler Datensatz nach Feature-Construction (Temporal-Sequencing)

Ergebnisse. Der Temporal-Sequencing-Parameter als Anteil der durch Präsens als Vordergrundinformation markierten Ereignisvorstellungen für die untersuchten obugrischen Texte zeigt (s. Plot 6.5.6) im Durchschnitt 26% aller Verben als *temporal-sequenced* (darin auch enthalten: `SUBPRED` mit `PRS`-Formen). Eingeschränkt auf die finiten Verben sind es 27%; schränkt man die Grundgesamtheit zusätzlich ein auf Verben auf der primären Handlungsebene, beträgt der Anteil von *temporal-sequenced*-Ereignisvorstellungen 35%.



Plot 6.5.6: Boxplot des Feature-Sets (Temporal-Sequencing)



Plot 6.5.7: Scatterplot nach GENRE-Klassen (Temporal-Sequencing)

Das heißt: Ein Viertel aller Ereignisvorstellungen und ein Drittel der Ereignisvorstellungen auf der primären Handlungsebene sind im Durchschnitt *temporal-sequenced* und damit (potentiell) als zentrale Foreground-Information markiert. Diesem Wert gegenüberzustellen ist der im folgenden Abschnitt 6.5.5 besprochene Wert des Subordinations-Backgroundings mit ca. 14%; es gibt also im Korpus im Durchschnitt doppelt so viel Vordergrund- wie Hintergrund-Markierung, was auch mit der Erwartung an narrative TWM übereinstimmt, sich auf die Haupthandlung als Vordergrund-Information zu konzentrieren und wenig Hintergrundinformationen zu geben (vgl. 3.8.4; s. auch Schulze 2019: 28).

Eine Auswertung bzgl. der Verteilung auf die Genres (s. Plots N.3f.) ergibt, dass sich die beiden journalistischen Texte im mittleren bis oberen Bereich bzgl. des Temporal-Sequencing-Anteils befinden und die Fate Songs in der oberen Hälfte; die ethnographischen Texte sind über den Wertebereich relativ stark gestreut. Bei den nichtnarrativen Texten (etwa `jou` als Reiseberichte, `fas` als Beschreibung von persönlichen Empfindungen) kann man von dem typischen Gebrauch der Präsensmarkierung im Obugrischen neben der Markierung von „narrative past“ ausgehen: „The Non-Past refers to the moment of speech, expresses the universal situation, or (immediate) future, for example: *man-l-ə-m* ’I am going, I (usually) go, I will go’ (go + NPAST + EP + 1SG)“ (Nikolaeva 1999: 26). Die relative Frequenz von Temporal-Sequencing-Formen in narrativen Texten bewegt sich – wie auch die anderen Angaben hier bezogen auf alle Verben in einem Text – durchschnittlich in einem Bereich von 25%.

Clustering (Plot N.1). Im Clustering (s. Plot N.2) zeigt sich eine narrative Subgruppe (3) längerer Erzählungen im Bereich über 20% Temporal-Sequencing, eine weitere narrative Subgruppe (2) der `anim_tal`-Texte um die Cranberry-Variationen mit niedrigem Temporal-Sequencing-Wert und eine heterogene Restgruppe (1) (u. a. auch die journalistischen Texte) mit höheren Werten. Eine kleine Clustergruppe (4) mit einem `fas`- und einem `eth`-Text liegen im Wertebereich noch darüber.

Klassifikation. Für das Temporal-Sequencing-Feature-Set erreicht keine Kategorisierung einen Kappa-Wert > 0.4.

6.5.5 Komplexitätsverlauf (Backgrounding)

- **Beobachtungsgegenstand**: verbale Einheiten
- **Feature-Attribut**: Region + Subordinationsstärke bzw. -typ
- **Normierung der Features**: relative Regionen-Frequenz
- **Skalierung des Feature-Sets**: keine (gemeinsame *bag*-Skala)

Die informationelle Vordergrund-Hintergrund-Strukturierung wird hier über Backgrounding durch Subordination modelliert, also über die Stärke und den Verlauf von Subordination in einem regionalen Bag-of-Tags-Modell. Für diese Auswertung der Subordinationsstruktur als Ausdruck des textuellen Komplexitätsverlaufs (s. 3.2.4; 3.8.4) werden zwei Modelle mit jeweils unterschiedlicher Regionenanzahl untersucht:

- ein einfaches Subordinationsmodell der **Subordinationsstärke** (`SUBPRED`) mit binärem Wertebereich (subordiniert: +/-)
- ein nach folgenden **Subordinationstypen** als Merkmalen differenziertes Modell (`SUBORD`):
  - `compC` = Komplementsatz
  - `subC` = subordinierter Clause
  - `ptcpVP` = nicht-clausewertige Partizipialkonstruktion (attributiv)

Feature-Construction und -Extraction. Datengrundlage für das Feature-Set des regional differenzierten Komplexitätsverlaufs ist (wie bei `TEMP_SEQ`) der relationale Primärdatensatz. Eine explizite Feature-Construction im Sinne einer Generierung neuer Merkmale ist hier nicht notwendig, da die vorhandenen Ereignis-bezogenen Merkmale der syntaktischen Funktion (`SYN`) und des Phrasentyps (`PHR`) bereits geeignete Merkmale für die Extraktion des Feature-Sets darstellen (s. Report 6.5.11).

	TEMP_SEQ	INFOSPEECH	SEM_DOM	PHR	SYN	TEMP_PRED	Text-ID
1	0	None	PERCEPTION	compC	SUBPRED		728
2	0	None	MOTION	finVP	PRED		728
3	0	None	PERCEPTION	compC	SUBPRED		728
4	0	None	MOTION	subC	SUBPRED		728
5	0	None	ACTION&PROCESS	finVP	PRED		728
6	1	None	PERCEPTION	finVP	PRED	PRS	728

Report 6.5.11: Relationaler Datensatz nach Feature-Construction (Komplexitätsverlauf)

Text-ID	1_SUBPRED	2_SUBPRED
728	0.18	0.11
730	0.27	0.03
732	0.13	0.20
741	0.00	0.00
742	0.12	0.04
750	0.10	0.05

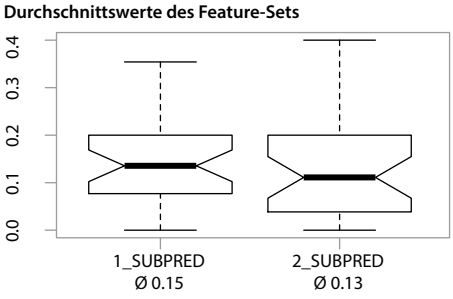
Report 6.5.12: Feature-Set für 2 Regionen (Subordinationsstärke)

Basierend auf diesen Merkmalen wird nun ein regional differenziertes Feature-Set mit zwei Regionen erstellt (Report 6.5.12), indem eine entsprechende Textpartitionierung vorgenommen und die relative Häufigkeit der subordinierten Einheiten (`SUBPRED`) pro Textregion als Merkmal berechnet wird.

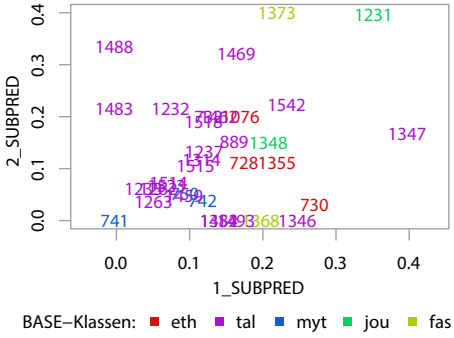
Für das `SUBORD`-Feature-Set (Report 6.5.13) werden zusätzlich die Phrasentyp-Informationen der `SUBPRED`-Einheiten berücksichtigt (d. h. mit in das Feature-Label aufgenommen).

Text-ID	1_compC	1_ptcpVP	1_subC	2_compC	2_ptcpVP	2_subC
728	0.12	0.00	0.06	0.11	0.00	0.00
730	0.03	0.03	0.21	0.00	0.00	0.03
732	0.03	0.03	0.08	0.07	0.03	0.10
741	0.00	0.00	0.00	0.00	0.00	0.00
742	0.04	0.08	0.00	0.00	0.08	0.04
750	0.05	0.03	0.02	0.01	0.03	0.02

Report 6.5.13: Feature-Set für 2 Regionen (Subordinationstypen)



Plot 6.5.8: Boxplot des Feature-Sets (Subordinationsstärke)



Plot 6.5.9: Scatterplot nach BASE-Klassen (Subordinationsstärke)

Ergebnisse. Die regional differenzierte Komplexitätsstärke der untersuchten obugrischen Texte zeigt eine geringe Anzahl an subordinierten verbalen Einheiten, wobei es hier keinen bedeutenden Unterschied zwischen den Regionenhälften der Texte

gibt (15% vs. 13%, s. Plot 6.5.8)[42], und entspricht damit der erwartbaren Prototypik für Volkserzählungen eines niedrigen Grads an Rekurrenz auf Hintergrundwissen (vgl. Schulze 2019: 28; Schulze 2004a: 555).

Clustering (Plots O.1 und O.3). Im Clustering des in zwei Regionen differenzierten `SUBPRED`-Feature-Sets wird (s. Plot O.2) eine kleine Gruppe von Texten mit erhöhter Komplexität abgetrennt (dazu gehört ein `fas`- und ein `jou`-Text). Außerdem lässt sich in den Gruppen keine einheitliche Tendenz einer Zu- oder Abnahme der Komplexitätsstärke zwischen erster und zweiter Texthälfte feststellen; d. h. bei der Komplexität (Subordinationsstärke) spielt – im Gegensatz zur Topik-Einführung – der zeitliche Verlauf im Text (hier über Regionen operationalisiert) keine Rolle.

Allerdings zeigt die Hopkins-Statistik, dass das 2-Regionen-`SUBORD`-Feature-Set mit Differenzierung nach Subordinationsstyp eine bessere Cluster-Tendency aufweist. Folgende Clustertypologie ergibt sich hier (s. Plot O.4):

- Gruppe 1 beinhaltet primär `magic_tal`-, `myt`-, und `eth`-Texte; diese größte Gruppe hat durchschnittliche Komplexität und eine Schwäche im `ptcpVP`-Bereich.
- Gruppe 2 beinhaltet beide `jou`-, einen `fas`- und einen `eth`-Text, also nichtnarrative Texte; stark im `ptcpVP`-Bereich, insgesamt leicht komplexer als Gruppe 1-Texte.
- Gruppe 3 beinhaltet `anim_tal`-Texte; stark im `compC`-Bereich (Komplementsätze) in der ersten Texthälfte, schwach im `ptcpVP`-Bereich.
- Ausreißer (4) ist ein Fate Song, sehr stark im `ptcpVP`-Bereich.

Klassifikation. Für die Komplexitätsverlauf-Feature-Sets erreicht keine Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*). Der Komplexitätsverlauf ist also, zumindest in der hier gewählten Operationalisierung über ein regional differenziertes Feature-Set, kein guter Prädikator für die verschiedenen Textsortenklassifizierungen.

6.5.6 Diskursstrukturelle Sequenzen

- **Beobachtungsgegenstand**: verbale Einheiten
- **Sequenzdaten**: *tag*-Folge des direkte-Rede-Status von Clauses
- **Normierung**: *maxlength* (textlängennormalisierte Sequenz)
- **Distanzmaß**: OM, euklidisch
- **Agglomerations-Methode**: Ward (Minimum Variance)
- **Klassifikationsmethode**: SVM, knn

42 Im Gegensatz zu der Clause-Komplexität oben (`CL_COMPLEX` = durchschnittlicher Anteil an Subordinationsmarkern pro Clause) wird hier die Subordinationskomplexität verbbasiert operationalisiert, also der Anteil an subordinierten verbalen Einheiten berechnet, und zwar sowohl Text- als auch Regionen-bezogen.

Das Diskurslayering (vgl. 3.8.5) als der Wechsel zwischen Haupthandlungsebene und eingelagerten dialogischen Diskursen wird hier zunächst als kategoriale *tag*-Folge modelliert.

Feature-Construction. Auf Grundlage des im Korpus entsprechend ausgezeichneten Status von Clauses (vgl. 5.3.5) kann das Vorkommen von prädikativen Einheiten in Passagen direkter Rede im relationalen Primärdatensatz bestimmt werden (`INFOSPEECH`-Merkmal in Report 6.5.14).

	TEMP_SEQ	INFOSPEECH	SEM_DOM	PHR	SYN	TEMP_PRED	Text-ID
1	0	None	PERCEPTION	compC	SUBPRED		728
2	0	None	MOTION	finVP	PRED		728
3	0	None	PERCEPTION	compC	SUBPRED		728
4	0	None	MOTION	subC	SUBPRED		728
5	0	None	ACTION&PROCESS	finVP	PRED		728
6	1	None	PERCEPTION	finVP	PRED	PRS	728

Report 6.5.14: Relationaler Datensatz nach Feature-Construction (Diskursstrukturelle Sequenzen)

Sequenz-Extraktion. Basierend auf diesen binären Präsenz-Werten des direkte-Rede-Status prädikativer Einheiten, werden textweise die globalen Abfolgen dieser binären *tags* (`SPEECH` vs. `NONE`) als sequentielle Modellierung der diskursiven Textstrukturierung extrahiert (Auflistung 6.34).

```
NONE-NONE-SPEECH-NONE-SPEECH-SPEECH-NONE-SPEECH-NONE-...
```

Auflistung 6.34: Zustandsfolge des textinternen Diskursstatus, Beginn Text 1484 (Diskursstrukturelle Sequenzen)

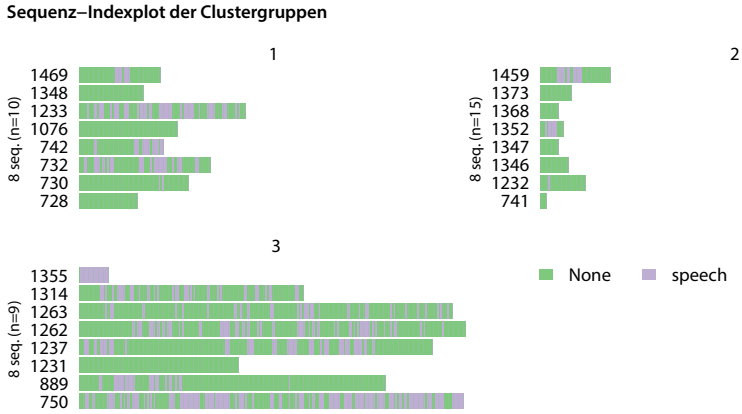
Zur Klassifizierung wird anschließend die `seqdist`-Funktion des `TraMineR`-Pakets verwendet, um auf diesen Folgen eine Optimal-Matching-Distanzmatrix zu erstellen.

Ergebnisse. Der gruppierte Index-Plot der Sequenz-Repräsentationen der textinternen Diskursstrukturierung der untersuchten obugrischen Texte (s. Plot P.3) zeigt, wie zu erwarten, dass narrative Texte (inkl. der mythologischen) durch Dialog-Passagen ausgezeichnet sind (so wie auch einige der ethnographischen Texte). In den kurzen Fabel-Erzählungen (`anim_tal`) findet sich meist eine Verdichtung von Dialogblöcken in der Mitte des Texts, bei längeren Erzählungen relativ gleichmäßig über den Text verteilt. Dagegen fehlen bei den Fate Songs und den journalistischen Texten solche Dialog-Passagen.

Clustering (Plot P.1). Im Clustering mit Ward-Agglomerationsmaß zeigen sich drei Hauptcluster (s. Plot 6.5.10):

- Gruppe 2 enthält vor allem die `anim_tal`-Texte (und `fas`); kurze Texte mit wenig oder keinem Dialog.
- Gruppe 3 enthält primär `magic_tal`-Texte und einen `jou`-Text; unterschiedliche Diskursstrukturierungen; vor allem längere Texte mit vielen Dialogpassagen.
- Gruppe 1 enthält `eth`-, `magic_tal`- und einen `jou`-Text; unterschiedliche Diskursstrukturierungen; vor allem Texte mittlerer Länge.

Allerdings sind in dieser Operationalisierung die Texte ohne Dialogpassagen (vor allem `jou`- und `fas`-Texte) über die Cluster verteilt; hier zeigt sich, dass die Operationalisierung über kategoriale *tag*-Sequenzen mit OM-Distanzmaß und *maxlength*-Normierung nicht gut geeignet ist, makrostrukturelle Verlaufstendenzen (hier von eingebetteten textinternen Diskursstrukturen) von Texten unterschiedlicher Länge zu modellieren. Das folgende Time-Warping der DTW-Modellierung erreicht hier ein besseres Ergebnis, indem es in der Lage ist, die Texte ohne Dialogpassagen als eigene Clustergruppe zu identifizieren.



Plot 6.5.10: Sequenz-Indexplot für Clustergruppen (Diskursstrukturelle Sequenzen)

Klassifikation. Zur Vorbereitung des Inputs für die Spectrum-SVM-Klassifikation werden die Kategorien der *tag*-Folgen (`None`, `speech`) mit ASCII-Zeichen der Länge 1 ersetzt; anschließend wird auf diesen *tag*-Folgen (1 Zeichen = 1 Zustand) mit 3-facher Kreuzvalidierung der Spectrum-SVM-Klassifikator trainiert und die Modell-Accuracy ausgegeben. In dieser Spectrum-SVM-Klassifikation erreichen die Sequenzdaten zur direkten Rede allerdings für keine Kategorisierung die Schwelle von Kappa > 0.4 (*moderate agreement*).

6.5.7 Diskursstrukturelle Partitur-Folgen

- **Beobachtungsgegenstand**: verbale Einheiten
- **Sequenzdaten**: numerische Folge des direkte-Rede-Status von Clauses

- **Normierung**: DTW
- **Distanzmaß**: dtwomitNA
- **Agglomerations-Methode**: complete
- **Klassifikationsmethode**: knn

Als Alternative zur kategorialen Sequenzanalyse von Diskurslayering wird nun eine Operationalisierung dieses informationsstrukturellen TWM-Parameters über Partitur-Folgen erprobt (vgl. 3.1.2 und 4.4.1.2). Als solche numerisch kodierten Diskursstruktur-Folgen können diese dann durch die Berechnung von Dynamic-Time-Warping-Distanzen analysiert werden. Dabei werden zwei Varianten von Partitur-Folgen untersucht, einmal als einfache Folge binär kodierter numerischer Präsenz-Werte des diskursiven Einbettungsstatus von Ereignisvorstellungen, außerdem als aggregierte Partitur-Folge der Vorkommenshäufigkeiten solcher eingebetteter Einheiten in der textweiten Folge von Sätzen als übergeordnete syntagmatische Einheiten.

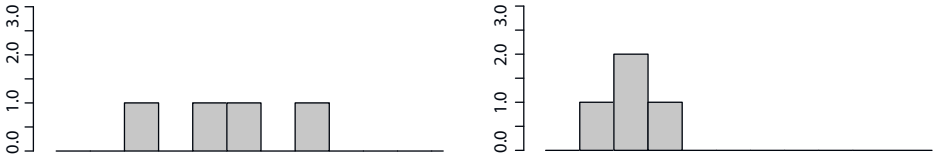


Abbildung 6.5: Binäre vs. aggregierte Partitur-Kodierung diskursiver Einbettung (Text 1484 „Cranberry")

Sequenzextraktion. Für die Erzeugung der einfachen, binär kodierten Partitur-Folgen können die zuvor berechneten kategorialen Folgen des direkte-Rede-Status prädikativer Einheiten (vgl. Auflistung 6.34) in numerische Folgen umgewandelt werden, indem die Werte des binären `INFOSPEECH`-Merkmals durch Kodierung mit Dummy-Variablen auf numerische Werte abgebildet werden:[43]

- `NONE` → `0` (Ebene der Haupthandlung, keine Einbettung)
- `SPEECH` → `1` (eingebetteter Dialog)

Die dabei erzeugten, numerisch kodierten Wertefolgen des Status direkter Rede in den Texten des Korpus werden textweise als Listen extrahiert (s. Auflistung 6.35).

```
0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0
```

Auflistung 6.35: Binäre Partitur-Folge, Text 1484 (Binäre DTW-Diskurspartitur)

Auf diesen numerischen Wertefolgen wird anschließend eine DTW-Distanzmatrix generiert, indem paarweise der Dynamic-Time-Warping-Abstand zwischen den numerisch-sequentiellen Repräsentationen der direkte-Rede-Struktur der Texte

43 Diese numerische Dummy-Rekodierung eignet sich nur für bestimmte (insbesondere binäre) kategoriale Variablen, vgl. James u. a. 2017: 130: „in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response […]."

berechnet wird (dazu wird das R-Paket `dtw` verwendet, vgl. 6.1.3.3). Für die Anwendung des Dynamic-Time-Warping-Algorithmus können die Folgen unterschiedlicher Länge sein (s. 4.4.1.2; vgl. Xing & Pei & Keogh 2010: 42) – eine Berücksichtigung der Textlänge wird hier also direkt in der Berechnung der DTW-Abstände zwischen den Wertfolgen vorgenommen.

Clustering (Plot Q.1). Im Clustering der binären Partitur-Kodierung zeigt sich (neben Gruppe 4 mit dem ethnographischen Text 1355, der überwiegend in direkter Rede gehalten ist, als Ausreißer) eine Einteilung der Texte in drei Hauptcluster, die in Plots 6.5.11ff. über die Barycenter-Durchschnittskurven der Clustergruppen visualisiert sind:

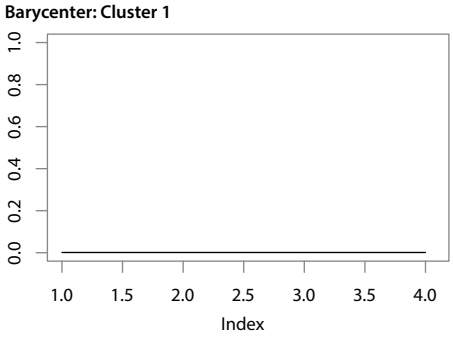
- Gruppe 1 enthält als Peripherie die Texte ohne dialogische Elemente (`jou`, `fas`, `eth` und wenige sehr kurze narrative Texte ohne direkte Rede).
- Gruppe 2 enthält als Kern die kürzeren narrativen Texte mit **Dialog-Mittelteil** (`anim_tal`; für eine detailliertere Auswertung s. 6.7.2; vgl. auch die Ergebnisse der Switch-Reference-Auswertung in 6.5.2).
- Gruppe 3 enthält im Sinne eines narrativen Subgenres die dialoglastigen längeren Erzählungen (alle Surgut-Khanty- und Mansi-Zaubermärchen `magic_tal` sowie die nordmansische mythologische Sage 750, die ein komplexes `SPEECH`-Muster zeigt).[44]

Es lassen sich in dieser DTW-Clustering-Analyse bzgl. der Abfolge diskursiver Einbettung also zwei narrative Subtypen im Korpus unterscheiden, nämlich längere, ‚klassische' Zaubermärchen mit stark dialogisch geprägtem Charakter (s. Plot 6.5.13) sowie eher kürzere Fabel- und Zaubermärchen mit einem Block direkter Rede nach kurzer Situationsexposition, der sich mit Unterbrechungen z. T. bis zum Ende des Haupthandlungsblocks hinzieht (s. Plot 6.5.12).

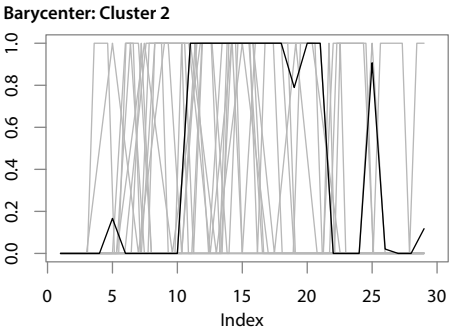
Indem die DTW-Analyse diese zwei dialoggeprägten narrativen Subgruppen (Cluster 2 und 3) von einer peripheren Gruppe (1) primär nicht-narrativer Genres (`fas`, `jou`, `eth`) ohne Dialog erfolgreich abtrennt, zeigt sich diese Analyse numerischer Sequenz-Repräsentationen der textinternen Diskursstrukturierung – im Gegensatz zu der OM-Sequenz-Analyse in 6.5.6, in der die Texte ohne Dialoge nicht als eigener Cluster abgetrennt werden – in der Lage, eine klare Trennung zwischen Texten ohne und Texten mit Dialogpassagen zu erreichen, und erscheint damit besser in der Lage, die für Volkserzählungen erwartbare Prototypik bzgl. dieses Parameters abzubilden, nämlich das Auftreten dialogischer Sequenzen in der

44 Die Yugan-Khanty-Zaubermärchen, die weniger komplex im Dialogbereich sind, werden dagegen mit den Tiermärchen in Clustergruppe 2 gruppiert.

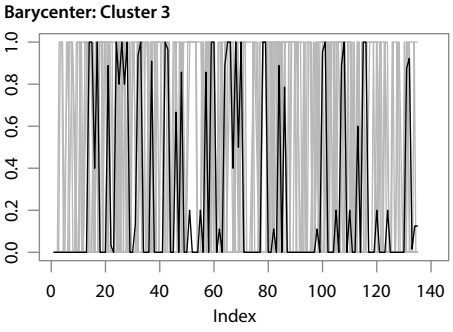
Interaktion zwischen den Protagonisten, in denen Ereignisse der Haupthandlung antizipiert oder reflektiert werden (s. Abschnitt 3.8.5; vgl. Schulze 2018: 203).



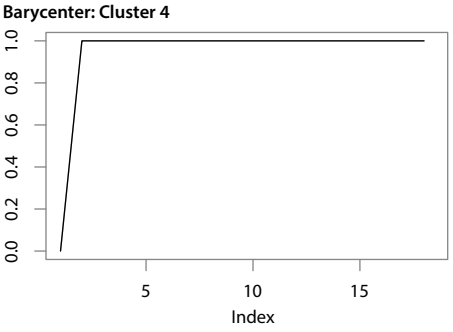
Plot 6.5.11: Barycenter Cluster 1 (Binäre DTW-Diskurspartitur)



Plot 6.5.12: Barycenter Cluster 2 (Binäre DTW-Diskurspartitur)



Plot 6.5.13: Barycenter Cluster 3 (Binäre DTW-Diskurspartitur)



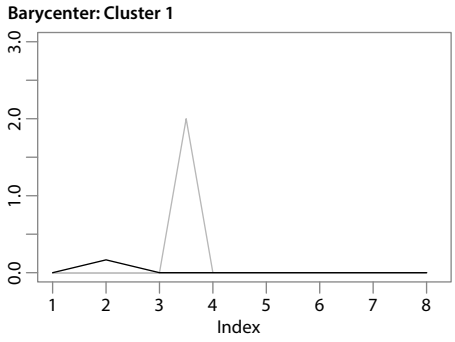
Plot 6.5.14: Barycenter Cluster 4 (Binäre DTW-Diskurspartitur)

Sequenzextraktion aggregierter Partitur-Folgen. Die Feature-Construction für die Operationalisierung über aggregierte Partitur-Folgen besteht in einer satzbezogenen Aggregation im Sinne einer Frequenzbestimmung der entsprechenden prädikativen Einheiten (`INFOSPEECH`) pro Satz als übergeordnete Informationseinheit, basierend auf Satz-ID-Angaben im relationalen Primärdatensatz (vgl. Report 6.5.14). Aus diesen Frequenzdaten der Anzahl von diskursiv eingebetteten verbalen Elementen pro Satz kann dann für jeden Text die aggregierte Partitur-Folge diskursstruktureller Einbettung als Satzfolge der jeweiligen Frequenzen erzeugt werden (s. Auflistung 6.36).

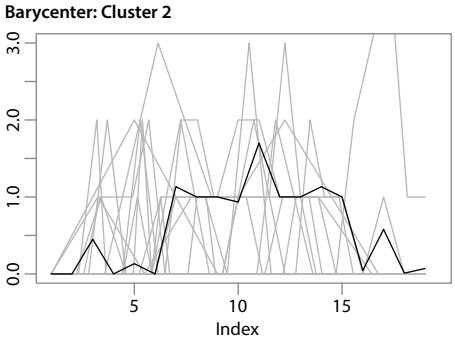
```
0, 1, 2, 1, 0, 0, 0, 0, 0, 0
```

Auflistung 6.36: Numerische Partitur-Folge, Text 1484 (Aggregierte DTW-Diskurspartitur)

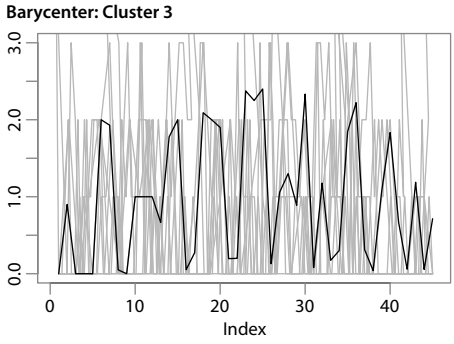
Clustering aggregierter Partitur-Folgen (Plot Q.2). Die aggregierten Partitur-Folgen erzielen ein vergleichbares Ergebnis im Clustering wie die einfachen Partitur-Folgen (s. die entsprechenden Barycenter-Plots der Clustergruppen in 6.5.15ff.). Die Gruppen 1 ohne SPEECH sind nahezu deckungsgleich; allerdings wird hier eine der narrativen Kerngruppen, nämlich die Gruppe 2 mit kurzem Dialogmittelteil, enger mit dieser Gruppe (1) ohne direkte Rede geclustert, sie wird also offensichtlich durch die satzbezogene Aggregation als weniger dialoglastig eingeschätzt.



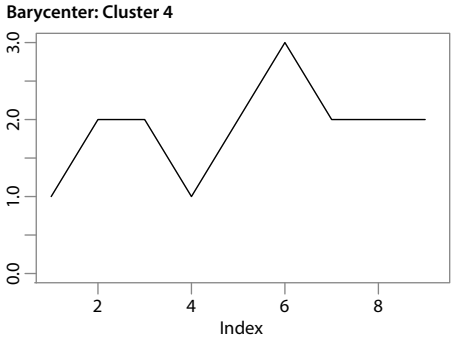
Plot 6.5.15: Barycenter Cluster 1 (Aggregierte DTW-Diskurspartitur)



Plot 6.5.16: Barycenter Cluster 2 (Aggregierte DTW-Diskurspartitur)



Plot 6.5.17: Barycenter Cluster 3 (Aggregierte DTW-Diskurspartitur)



Plot 6.5.18: Barycenter Cluster 4 (Aggregierte DTW-Diskurspartitur)

Klassifikation. Für eine Anwendung der knn-Klassifikation auf DTW-Distanzdaten (vgl. Giorgino 2009: 3) wird die DTW-Distanzmatrix verwendet, um die k Nächsten-Nachbarn eines Textes zu errechnen und die Klasse zu bestimmen, in der die meisten der Nachbarn liegen (vgl. 4.4.3). Die knn-Klassifikation, die hier nur testweise ohne Kreuzvalidierung durchgeführt wurde, zeigt bei der binären COMM_SIT-Klassifizie-

rung die höchste Accuracy (0.82); `BASE` und `GENRE` erreichen hier nur eine schlechte Accuracy; die aggregierten Partitur-Folgen schneiden ähnlich ab.

6.6 Gesamtbewertung der Parameter

6.6.1 Gesamtmodell der Feature-basierten Parameter

Bisher wurden im Rahmen des Forschungsvorhabens einer TWM-Operationalisierung und -Rekonstruktion durch automatische Klassifizierungsmethoden solche Feature-basierten Modellierungen untersucht, die entweder Operationalisierungen einzelner TWM-Parameter waren oder die, wenn mehrere Parameter in einem Feature-Set kombiniert wurden, zumindest auf einzelne Strukturierungsbereiche textuell kodierter Modelle beschränkt waren. Für eine Modellierung von TWM als kognitive, quantitativ-schematische Weltmodelle des global-strukturellen, referentiellen, relationalen und informationsstrukturellen Aufbaus genrespezifischer Text-Modelle sowie auch für eine datengestützte Gewichtung der verschiedenen untersuchten Merkmale bzgl. ihrer Relevanz als TWM-Parameter für die Differenzierung von Texten verschiedener Genres (vgl. Schulze 2019: 31; Schulze 2020: 628f.) werden in der folgenden Auswertung die Feature-basierten Parameter in einem **Gesamt-Feature-Set** kombiniert. In einer hierarchischen Clusteranalyse kann so aus den obugrischen Korpusdaten eine induktive textstrukturelle Clustertypologie gewonnen werden, die man als potentiellen Ausdruck verschiedener TWM verstehen kann, die sich als kognitive Strukturmodelle in der Textproduktion der Sprecher niederschlagen und entsprechend über die hier erprobte Methodik aus dem Sprachgebrauch rekonstruieren lassen.

Anschließend erfolgt eine Analyse der Relevanz der einzelnen Merkmale für diese sprachgebrauchsbasierte Clustertypologie. Dabei wird eine datengestützte **quantitative Gewichtung**[45] der Parameter vorgenommen, um ihre Bedeutung für die Unterscheidung der induktiv im Korpus entdeckten Strukturtypen zu bestimmen.[46] Hierfür werden mithilfe des Random-Forest-Verfahrens die Feature-Importance-Werte für die Clustergruppen des Feature-basierten Gesamtmodells berechnet; durch die Analyse dieser Werte kann untersucht werden, welche textstrukturellen Merkmalsdi-

45 Im Gegensatz zur bisherigen hypothesenbasierten Auswahl der Merkmale für die Feature-Sets der einzelnen Parameter (vgl. 4.1.2.2), die sich an der bei Schulze (2020) aufgestellten Systematik quantitativer TWM-Parameter bzgl. kognitiver Domänen orientierte (vgl. 3.1), ist so eine datengestützte Feature-Selection als Rekonstruktion der relevanten TWM-Parameter möglich.

46 Von einer solchen quantitativen Gewichtung als Ranking der TWM-Parameter nach ihrer Bedeutsamkeit für die Genre-Diskrimination ist nach Schulze (2019: 31) eine inhaltliche Gewichtung dieser Parameter zu unterscheiden, die sich auf deren intensionale Bedeutung bezieht. Die in dieser Arbeit angewandten automatischen Klassifizierungs- und Mustererkennungsmethoden können ihrer Natur als statistisch-quantitative Modelle gemäß natürlich nur eine quantitative Gewichtung leisten und werden in der Ergebnisdiskussion in Abschnitt 6.7 interpretativ ergänzt.

mensionen Genres bzw. Subgenres im Korpus unterscheiden und welche Dimensionen dagegen peripher sind (s. Schulze 2019: 31; Schulze 2020: 592). Dieser Ansatz ermöglicht also die explorative Identifizierung der prototypischen Muster (Schulze 2019: 15), die die Texte eines Genres auszeichnen und die man in ihrer Gesamtheit als das deren Textproduktion und -rezeption zugrunde liegende, kognitive Strukturmodell (TWM) verstehen kann (s. Schulze 2019: 14f.). Auf diese Weise erhält man eine Beschreibung der quantitativen kognitiven Texttypen, die induktiv im Korpus identifiziert wurden, basierend auf den für ihre Differenzierung relevanten Merkmalsdimensionen.

Zusätzlich werden Feature-Importance-Werte für theoriegestützte, text- und diskurslinguistisch begründete Apriori-Textkategorisierungen berechnet (vgl. 5.2.4); die Ergebnisse, also die Relevanz der angenommenen TWM-Parameter für die Kategorisierung, werden sowohl untereinander als auch mit den Clustergruppen verglichen.

Feature-Construction. Folgende, bereits zuvor für die Einzelmodelle konstruierte Merkmale werden im Gesamt-Feature-Set zusammengefasst:

- **Globale textstrukturelle Merkmale**
- **Referentielle Distanz und Topik-Persistenz**
- **Durchschnittlicher Topikalitätsquotient**
- **Referenten-bezogenes Topikalitätsmodell**
- **Ereignistypik**
- **Häufige Übergänge von Ereignistypen**
- **Kontrastiver Fokus**[47]
- **Regional differenzierte Topik-Einführungsstärke**
- **Temporal-Sequencing-Stärke** (Foregrounding)
- **Komplexitätsstärke**[48] (Backgrounding)

Feature-Extraction. Zur Homogenisierung des durch die Zusammenstellung verschiedener Features mit unterschiedlichen Skalen heterogenen Merkmalraums erfolgt zunächst eine **Skalierung** des Gesamt-Feature-Sets durch z-Standardisierung, um anschließend die etablierten Klassifizierungsmethoden anwenden zu können. Auf eine Anwendung von automatischen Feature-Subset-Selection-Methoden wie Recursive-Feature-Elimination (s. Guyon & Elisseeff 2006: 15) oder Analysen von Feature-Korrelationen zur Bestimmung der optimalen Anzahl und Auswahl von

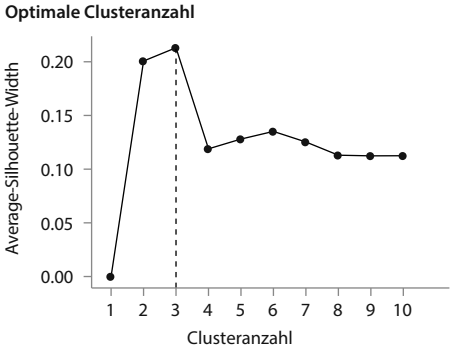
47 Die Auswahl der pragmatischen Features wird auf eine Fokusart eingeschränkt, nämlich `CTR` als kontrastive Fokussierung, die im Gegensatz zu den anderen hier annotierten Fokusarten im Obugrischen eindeutig grammatikalisch markiert ist (Pronomen statt Nullanapher).

48 An dieser Stelle wird eine textweite Variante der Stärke von subordinierten Einheiten verwendet.

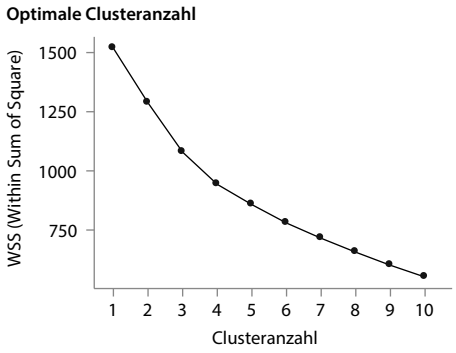
Merkmalen (s. Guyon & Elisseeff 2006: 12) wurde im Rahmen dieser explorativen Pretest-Studie verzichtet.

Ergebnisse. In der nachfolgenden Clusteranalyse des Gesamt-Feature-Sets zeigt sich, dass Varianten eines Märchens jeweils zusammen gruppiert werden (auch solche, die über lange Zeiträume und Dialektgrenzen getrennt aufgezeichnet wurden), ebenso wie die Zaubermärchen mit ähnlichem Handlungsstrukurtyp, s. dazu auch **Abbildung 6.6 mit markierten Subgenres und Varianten** in der Ergebnisdiskussion in Abschnitt 6.7. Gemeinsam bilden diese einen narrativen Hauptcluster (1), den man (evtl. zusammen mit seinem Schwestercluster (2), der aus Varianten des kurzen, fabelartigen `Cranberry`-Tiermärchens besteht) als den **Kerntyp narrativer Strukturierung** des obugrischen Korpus verstehen kann (vgl. 5.2.3). Die Texte mythologischen Inhalts sind über diese beiden narrativen Cluster (1 und 2) verstreut, sind also nicht als eigenständiges Genre zu identifizieren; auch die meisten ethnographischen Texte sind (als persönliche, nicht-fiktive Erzählungen) Teil des narrativen Hauptclusters, darin aber gemeinsam in einem Subcluster gruppiert.

Als **Peripheriegruppe** (3) bilden die Fate Songs zusammen mit dem Zeitungstext 1231 (`Faraway`, eine Reisebeschreibung) und dem ethnographischen Text 1355 (`Trap`, eine Wegbeschreibung) einen kleinen Cluster von Texten mit nicht-narrativer Struktur und untypischer Erzählperspektive (lyrisches *Ich/Es* in den Fate Songs, *Du* im ethnographischen Text, *Wir* im Zeitungstext), die insbesondere gekennzeichnet sind durch geringe Frequenz von Handlungsverben, geringe Topic-Continuity sowie eine hohe referentielle Informationsdichte.



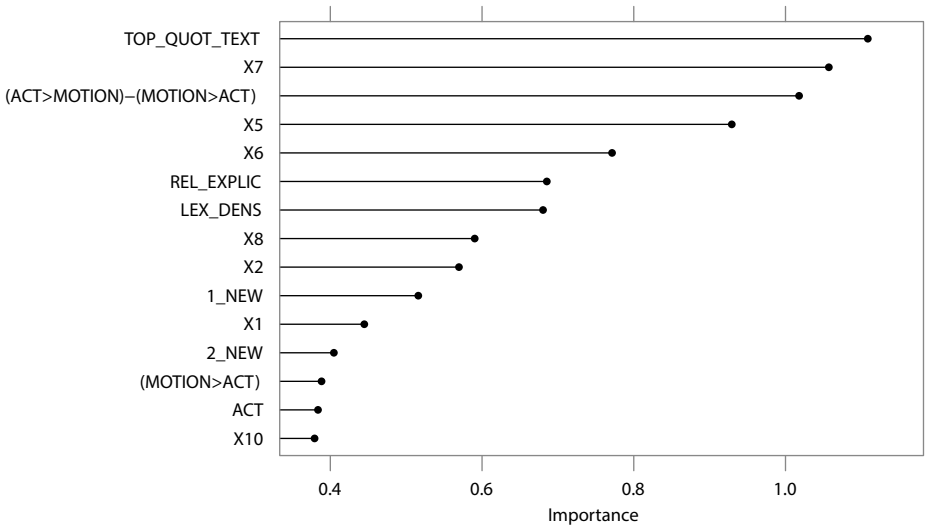
Plot 6.6.1: Silhouette-Plot (Gesamt-Feature-Set)



Plot 6.6.2: Elbow-Plot (Gesamt-Feature-Set)

Clustering (Plot R.1). Gemäß der im Silhouette-Plot 6.6.1 festgestellten optimalen Clusteranzahl von drei Clustern zeigen sich im Clustering des Gesamt-Feature-Sets zwei von einer **Hauptgruppe 1** (= rot) getrennte textstrukturelle Clustergruppen.

Berücksichtigt werden in der folgenden Clusteranalyse primär die Merkmale, die sich in der Feature-Importance-Analyse der Clustergruppierung als distinktiv zeigen (s. Plot 6.6.3). Die Analyse beginnt dabei mit den beiden peripheren Gruppen, um dann kontrastiv dazu die Textstruktur-Charakteristik der Hauptgruppe als **narrativer Kerngruppe** des obugrischen Korpus herauszuarbeiten (vgl. 5.2.3).



Plot 6.6.3: Feature-Importance Top 15 für Clustergruppen (Gesamt-Feature-Set)

Gruppe 2 (= grün) als narrative Subgruppe, die primär die Varianten des kurzen `Cranberry`-Fabelmärchens enthält, zeichnet sich aus durch eine **geringe Anzahl an Referenten** (d. h. einen hohen textweiten Topikalitätsquotienten, s. Plot 6.6.4; gleichzeitig wenig elaboriert und geringe referentielle Explizitheit, s. Plots 6.6.7f. und R.6f.) und einen **hohen Anteil an Handlungsverben** (s. Plot 6.6.5). Der relative Anteil der Einführung neuer Topiks (`1_NEW` in Plot 6.6.4) ist ähnlich wie bei der narrativen Hauptgruppe (1). Allerdings ist aufgrund der Konzentration der Handlung auf wenige Referenten in der sortierten Verteilung der Topikalitätsquotienten pro Referent ein **stärkerer Abfall deren Topikalität** festzustellen (s. Plot R.2; vgl. auch Cluster 3 in Plot 6.3.12 des Clusterings von Abschnitt 6.3.5, der ebenfalls primär die `Cranberry`-Texte enthält). Die kurzen Texte dieser narrativen Subgruppe unterscheiden sich von den anderen narrativen Texten insbesondere in der Reduktion des kognitiven Text-Modells auf die elementaren Informationen der Handlung und in der damit einhergehenden Beschränkung auf eine minimale Anzahl an Referenten, weiter durch fehlende oder minimale Lokalisierungen, da die Handlung in diesem Texttyp von Tiermärchen an einem nicht näher spezifizierten bzw. nur kurz zu Beginn eingeführten

Ort in einer kurzen Zeitspanne stattfindet. Damit bestätigt sich auch strukturell der in 5.2.3 aufgrund inhaltlicher Überlegungen postulierte Charakter dieser Subgruppe als **fabelartige Erzählungen**.[49] Deren skizzenhafte Handlungsfolge drückt sich in niedriger durchschnittlicher **referentieller Distanz** im LOC-Bereich aus (s. Plot 6.6.6 und R.3), was auch mit der regelmäßigen Verwendung des Passivs im Obugrischen zur Kodierung von Topikalitätsunterschieden und zur Aufrechterhaltung von Topikkontinuität zusammenhängt.[50]

Gruppe 3 (= blau) als periphere Gruppe nicht-narrativer Texte mit den beiden lyrischen Fate Songs (`fas`), dem Zeitungstext `Faraway` (1231) und dem ethnographischen Text `Trap` (1355) zeichnet sich im Vergleich zu den beiden narrativen Gruppen aus durch einen **geringen Anteil an Handlungsverben** (vgl. Plot 6.6.5). Stattdessen findet sich durchschnittlich ein höherer Anteil an Verben der Wahrnehmung (s. Plot R.4)[51] sowie eine **hohe relative Häufigkeit von Topikeinführungen** (s. Plot 6.6.4 und R.8) bei gleichzeitig ähnlich **hoher absoluter Anzahl an Referenten** im Text-Modell wie bei der Hauptgruppe 1 (entspricht einem niedrigen textweiten Topikalitätsquotienten, s. Plot 6.6.4; gleichzeitig stärker nominal elaboriert und höhere referentielle Explizitheit, vgl. Plots 6.6.7 und R.6f.). Anders als bei den Tiermärchen der Gruppe 2 gibt es hier somit im Text-Modell eine größere Anzahl nur schwach topikaler Referenten mit nur wenigen Referentenerwähnungen (Refe-

49 Vgl. Dithmar 1974: 1034: „Die Fabel bietet keine abgeschlossene Handlung, sondern gibt nur einen Ausschnitt. Der Vorhang auf der Bühne wird nur einmal kurz aufgezogen, um dem Zuschauer einen Einblick zu gewähren."

50 Im Fall der kurzen Fabelmärchen, in denen alle der wenigen Referenten relativ stark topikal sind, finden sich regelmäßig Hauptreferenten in adverbialer Position passiver Konstruktionen (LOC) – so etwa in den `Cranberry`-Texten die Pflanzenreferenten, ebenso aber auch das Feuer (das eigentliche Textthema), von dem diese verbrannt werden.

51 Fate Songs sind „lyric songs, where the moods, experiences and fate of the author (a man or a woman) are described" (Ojamaa & Ross 2004: 134); so z. B. in Text 1368 `GameVoicedSong`, Sätze 3–5:

(10) jæni-ŋ konɣæ-n kʷo:ntl-eɣ-m
big-PROPR land_neck-DLAT listen-PRS-1SG
I listen in the direction of the big portage route, (PM, 1368: 3)

(11) suj-iŋ ke:rpaip tæ sujt-i
sound-PROPR bell EMPH1 be_heard-PRS[3SG]
A resounding bell can be heard, (PM, 1368: 4)

(12) kommiŋ suj suj-æ suml-əs
ringing sound sound-SG<3SG ring_out-PST[3SG]
The sound of a ringing sound rings out. (PM, 1368: 5)

rententokens) pro Referententype[52] (s. 3.8.1; vgl. Biber 1992b: 232).[53] Entsprechend verteilt sich hier auch die Topikalität gleichmäßiger auf die dem Hauptreferenten als Textthema nachfolgenden Referenten (s. Plot R.2; vgl. auch Plot 6.3.11). Aufgrund der hohen relativen Häufigkeit von Topikeinführungen kann man annehmen, dass in dieser Clustergruppe die Topic-Continuity im Gegensatz zu den narrativen Texten entsprechend geringer ist, da mit der Einführung neuer Topiks gleichzeitig ein Referenzwechsel einhergeht, was auch das Clusteringresultat der geglätteten Switch-Reference-Struktur in Plot S.1 anhand der Clustergruppe 1 um die `jou-` und `fas`-Texte mit **hohem Anteil an Switch-Reference**-Phasen bestätigt (s. Plot 6.7.4; vgl. auch Plot L.3).

Hauptgruppe 1 (= rot) als der Kern der durch induktive Mustererkennung des Clusterings gefundenen Textstruktur-Typik zeichnet sich im Gegensatz zu den anderen Gruppen aus durch Übergänge der Art `ACTION > MOTION > ACTION` (s. Plot 6.6.5), also durch **Verkettungen von Bewegungs- und Handlungsereignissen**.[54] Ebenso gibt es u. a. eine höhere durchschnittliche relationale Explizitheit und eine höhere lexikalische Dichte (s. Plots 6.6.7f. und R.6f.). Die Analyse des textweiten Topikalitätsquotienten zeigt Werte deutlich unter denen der Fabelmärchen in Cluster 2 (d. h. eine **höhere Anzahl an Referenten**; s. Plots 6.6.4 und R.3); im Gegensatz zur Peripherie (3) ist die **relative Häufigkeit der Einführung neuer Topiks geringer** (s. Plot 6.6.4 und R.8), da hier (ebenso wie in dem anderen narrativen Cluster 2) kontinuierliche Topiks länger aufrechterhalten werden (vgl. die narrativen Gruppen 2 und 3 in Plot 6.7.4 des geglätteten Switch-Reference-Clusterings; vgl. auch Cooreman 1987: 14). In den Texten dieses primär narrativen Clusters liegt also ein höherer Grad an **Topic-Continuity** vor (entsprechend lässt sich eine höhere referentielle Inferenz feststellen, also ein höherer Anteil an Nullanaphern; s. Plot R.6). Die Topikalität der Referenten nimmt im Gegensatz zur Peripherie kontinuierlich ab (ca. 1/3 pro Referent[55]; s. Plot R.2), was sich für die Zaubermärchen als **Abstufung in der Topikalität der Referenten** gemäß ihrer Rollen als Protagonisten, Antagonisten

52 Dieses Token-Type-Verhältnis gilt auch für den regional differenzierten Parameter `1_NEW` (vgl. 3.8.1), da für die erste Textregion die Anzahl neuer Referenten der Anzahl an Referententypes entspricht (mit der Ausnahme bei Einführung mehrerer Referenten als Gruppe, die hier als eine Topik-Einführung zählt).

53 Biber 1992b: 232: „It is noteworthy that spot news, [...] and [...] academic prose are extremely 'informational' in both respects: they have the highest absolute frequencies of new referents, and proportionally they use very high percentages of their referring expressions for new references [...]."

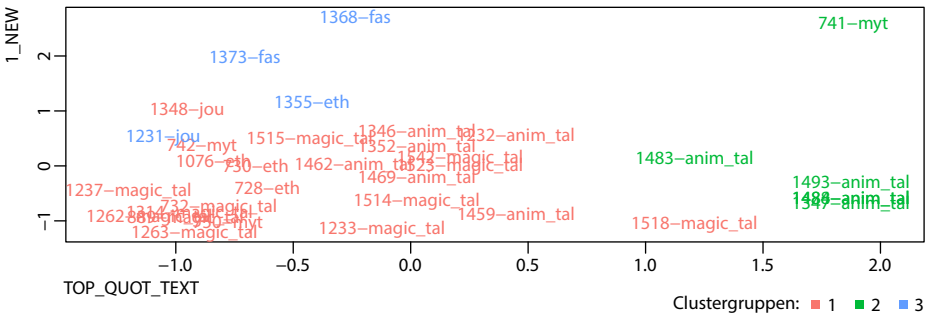
54 Dies kann dahingehend interpretiert werden, dass auf der kognitiven Karte dieser Texte mehr Positionsveränderungen stattfinden als in den kurzen Fabelmärchen des Clusters 2, dass diese *maps* also eine höhere Anzahl an *landmarks* umfassen (vgl. Abschnitt 3.5).

55 Dies ist der in 6.3.5 für das Korpus festgestellte Durchschnitt der Abnahme des Topikalitätsquotienten pro Referent, dem die Gruppe 1 entspricht (in dem skalierten Gesamt-Feature-Set liegen für Gruppe 1 die Werte der Topikalitätsverteilung annähernd bei 0, also dem zentrierten Mittelwert, s. Plot R.2).

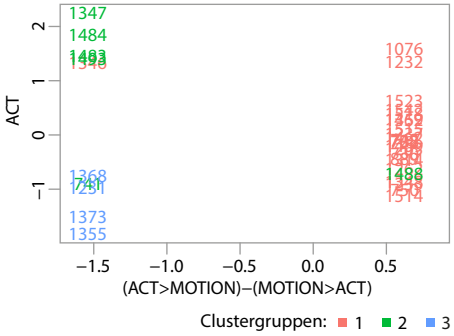
bzw. Nebenfiguren verstehen lässt (z. B. Helfer; vgl. das 7-Personenschema bei Propp 1972: 98).

Zusammenfassend kann man diese drei induktiv identifizierten informationsstrukturellen Texttypen über die in Plot 6.6.4 abgebildeten Dimensionen des textweiten **Topikalitätsquotienten** (als inverse *absolute* referentielle Informationsdichte) sowie der **Topik-Einführungsstärke** (als *relative* referentielle Informationsdichte) erfassen:

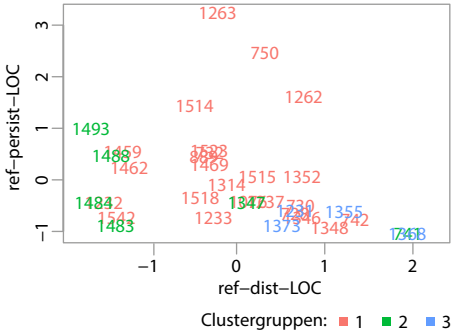
- **Cluster 2 (grün)** als Subgruppe kurzer Tiermärchen mit geringer absoluter Anzahl an Referenten (hoher Topikalitätsquotient, **niedrige absolute Informationsdichte**), die aber wiederholt aufgenommen werden (mehrere stark topikale Hauptreferenten, vgl. Plot R.2, vgl. auch Cluster 3 in Plot 6.3.12); der Cluster weist entsprechend eine **niedrige relative Informationsdichte** auf.
- **Cluster 3 (blau)** als Peripherie nicht-narrativer, informativer Texte mit hoher absoluter Anzahl an Referenten (**hohe absolute Informationsdichte**), die aber nur schwach wiederholt werden (viele Neueinführungen, also viele, eher schwach topikale Referenten), d. h. der Cluster weist eine **hohe relative referentielle Informationsdichte** auf (ein vergleichsweise hoher Anteil von Referentenerwähnungen sind Einführungen neuer Topiks; vgl. Plot R.8).
- **Cluster 1 (rot)** als narrativer Kern mit einer verhältnismäßig hohen absoluten Anzahl an Referenten (**hohe absolute Informationsdichte**), die aber wiederholt aufgenommen werden, was mit einer **niedrigen relativen Informationsdichte** im referentiellen Bereich einhergeht (vgl. Biber 1992b: 231f.).



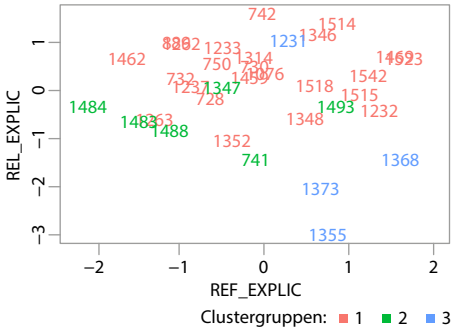
Plot 6.6.4: Textweite Topikalitätsstärke (TOP_QUOT_TEXT) und relative Häufigkeit von Topik-Einführungen in der ersten Texthälfte (1_NEW) als Kennzeichen (inverser) absoluter bzw. relativer referentieller Informationsdichte nach Clustergruppen (Gesamt-Feature-Set)



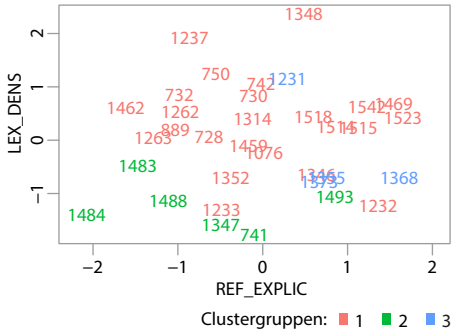
Plot 6.6.5: Handlungsbezogene Features nach Clustergruppen (Gesamt-Feature-Set)



Plot 6.6.6: Referentielle Distanz und Topik-Persistenz im LOC-Bereich nach Clustergruppen (Gesamt-Feature-Set)



Plot 6.6.7: Referentielle und relationale Explizitheit nach Clustergruppen (Gesamt-Feature-Set)

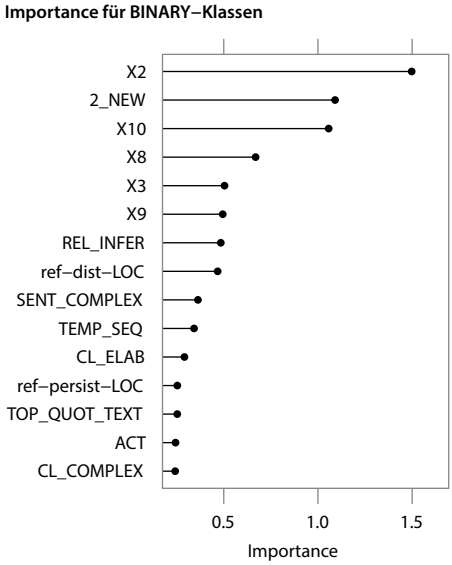


Plot 6.6.8: Referentielle Explizitheit und lexikalische Dichte nach Clustergruppen (Gesamt-Feature-Set)

Klassifikation. In der Klassifikation mit Random-Forest erreicht das Gesamt-Feature-Set für die `COMM_SIT`-Kategorisierung einen Kappa-Wert > 0.6 (*substantial agreement*), für die `GENRE`- und `DISC_STRUCT`-Kategorisierung einen Kappa-Wert > 0.4 (*moderate agreement*); die `BASE`-Kategorisierung bleibt unter dieser Schwelle.

Eine für die Gesamtauswertung neu eingeführte, **binäre Kategorisierung** (`BINARY`; vgl. Biber 1992a), die auf der primären Textsorten-Einteilung (`BASE`) aufbaut und die deren narrative Klassen `tal` und `myt` als narrativen Kern des Korpus in einer Klasse zusammenfasst sowie die übrigen Klassen `jou`, `fas` und `eth` in einer Klasse nicht-narrativer Texte vereint, erreicht einen Kappa-Wert von 0.5 (*moderate agreement*).[56] Wichtigste Merkmale für diese `BINARY`-Kategorisierung (Plot 6.6.9) sind die **Topikalitätsstärke** des zweit- bzw. zehnthäufigsten Referenten sowie die relative referentielle Informationsdichte in der zweiten Texthälfte.

56 Die `BINARY`-Kategorisierung erreicht auch die höchste Accuracy mit 0.88.



Plot 6.6.9: Feature-Importance Top 15 für BINARY-Klassen (Gesamt-Feature-Set)

Für die beiden textsortenbezogenen Genre-Kategorisierungen (`BASE` und `GENRE`) zeigen die Feature-Importance-Werte ein relativ gleichmäßig absteigendes Ranking (s. Plots 6.6.10f.). **Topikbezogene Größen** wie die relative Häufigkeit von Topik-Einführungen, die referentielle Distanz und die Topik-Persistenz (insbesondere im LOC-Bereich) sowie der Topikalitätsquotient (textweit und pro Referent) spielen hier eine wichtige Rolle für die Differenzierung der Klassen. Allerdings ist für die `GENRE`-Kategorisierung, die eine Subdifferenzierung der narrativen Texte in längere Erzählungen (`magic_tal`) und kürzere Tiermärchen (`anim_tal`) vornimmt, zusätzlich die **Ereignistypik** im `MOTION`-Bereich differenzierend für diese Subgenres – im Gegensatz zu den Textsorten-Einteilungen `BASE` sowie `BINARY`, die jeweils einen narrativen Kern von einem oder mehreren peripheren Texttypen trennen und für diese Trennung primär Topikalitätskriterien priorisieren (vgl. Plot 6.6.9). Zusätzlich ist für die `GENRE`-Einteilung auch der **textweite Topikalitätsquotient** relevant, was mit dem Ergebnis der Clusteranalyse einer Differenzierung kurzer Fabelmärchen von längeren Erzählungen über deren absolute referentielle Informationsdichte übereinstimmt (s. Plot 6.6.4; vgl. auch Plot F.3).

Für die beiden diskursbezogenen Kategorisierungen ist jeweils eines der globalen Textstrukturmerkmale mit Abstand am wichtigsten (s. Plots 6.6.12f.): für die `COMM_SIT`-Kategorisierung die **Clause-Elaboration**; für die `DISC_STRUCT`-Kategorisierung die **referentielle Explizitheit**.

BASE	GENRE	DISC_STRUCT	COMM_SIT
0.20	0.18	0.10	-0.01

Report 6.6.1: Adjusted Rand-Index (Gesamt-Feature-Set)

Die im hierarchischen Clustering gefundene Klasseneinteilung (s. Plot 6.6.3) zeigt primär **topikalitätsbezogene** informationsstrukturelle Features als relevant und damit eine größere **Übereinstimmung** mit den beiden **textsortenbezogenen** als mit den beiden diskursbezogenen Einteilungen (vgl. auch Report 6.6.1). Darüber hinaus ist für diese induktive Clustertypologie die Abfolge der **Ereignisübergänge** `ACTION > MOTION > ACTION` eines der wichtigsten Unterscheidungsmerkmale, während dies für keine der vier Apriori-Kategorisierungen eine wesentliche Rolle spielt.

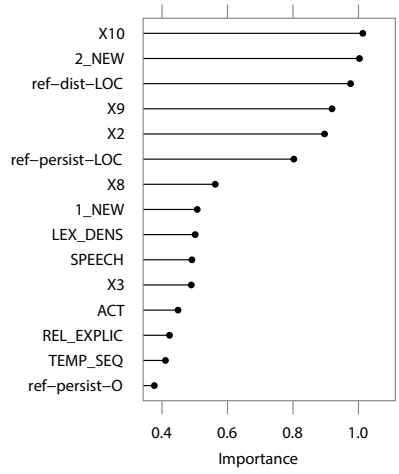
Bei der Klassifikation mit Random-Forest schneidet das textstrukturelle Gesamt-Feature-Set für die Vorhersage der verschiedenen Apriori-Textkategorisierungen besser ab als die zuvor untersuchten Teilmodelle kognitiver Strukturbereiche: Während diese größtenteils nur für die binäre COMM_SIT-Klassifizierung[57] einen Kappawert > 0.4 (*moderate agreement*) erreichen, erzielt dies das Gesamtmodell für alle bis auf die BASE-Klassifizierung. Bei den Einzelmodellen hingegen erreicht nur das Feature-Set der Topik-Einführungen (6.5.1) sowie das kombinierte Feature-Set globaler textstatistischer Merkmale (6.2.4) für andere als die COMM_SIT-Klassifizierung Kappa > 0.4 (in beiden Fällen für die DISC_STRUCT-Klassifizierung).

Betrachtet man abschließend die Klassifikationsresulate für die drei primär textintern ausgerichteten Genre-Einteilungen (BASE, GENRE und DISC_STRUCT; ohne die Kommunikationstypisierung COMM_SIT), die mit jeweils unterschiedlichen theoretischen Schwerpunkten narrative Texttypen von informativen und anderen nichtnarrativen Texttypen trennen (s. 5.2.3f.), so ergeben sich folgende Beobachtungen: Die Tatsache, dass die binäre Klassifizierung (BINARY) ein deutlich höheres Agreement (Kappa: 0.50) aufweist als die BASE-Klassifizierung (Kappa: 0.18), auf der diese basiert, kann man als Hinweis darauf sehen, dass – neben eines anzunehmenden positiven Einflusses der geringeren Klassenanzahl auf das Klassifikationsergebnis – das textstrukturelle Gesamt-Feature-Set grundsätzlich geeigneter ist, **narrative von nicht-narativen Texten zu unterscheiden** als die verschiedenen nicht-narrativen Textsorten der BASE-Klassifizierung zu trennen.[58] Daran anschließend zeigt das Agreement der GENRE-Klassifizierung (Kappa: 0.43), dass deren Subdifferenzierung der Textsorten im narrativen Bereich durch Aufsplitten der tal-Klasse in anim_tal und magic_tal die (relative) Accuracy erhöht, obwohl die Klassenzahl zunimmt. Dies spricht – die Clusteringergebnisse unterstützend – für das Vorliegen von zwei narrativen Subgenres im Datensatz mit unterschiedlichen TWM-Strukturschemata (Zaubermärchen vs. Fabel-Tiermärchen). Dass auch die DISC_STRUCT-Funktionseinteilung (mit den vier Klassen narrativ, prozedural, verhaltensbezogen und expositorisch) als strikt theoriegestützte, diskursfunktionale Genre-Typisierung nach den Kriterien von Longacre 1983 (vgl. 5.2.4) in der Klassifikation mit dem Gesamt-Feature-Set ein moderates Agreement erreicht (Kappa: 0.40), deutet darauf hin, dass die informations-, referenz- und handlungsstrukturellen Merkmale des

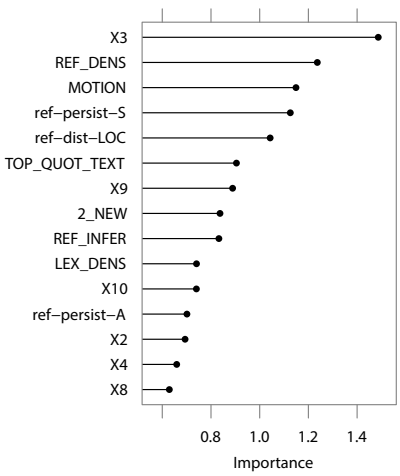
57 Dies auch aufgrund der geringeren Klassenanzahl im Vergleich mit den anderen Kategorisierungen, vgl. Sim & Wright 2005: 264: „The larger the number of scale categories, the greater the potential for disagreement, with the result that unweighted kappa will be lower with many categories than with few."

58 Diese nicht-narrativen Genres sind allerdings in dem vorliegenden begrenzten Pretest-Datensatz obugrischer Texte auch stark unterrepräsentiert, d. h. der Datensatz ist in dieser Hinsicht unbalanciert, da er primär Erzähltexte enthält, deren Struktur und mögliche Subtypen in dieser explorativen Studie vorrangiger Untersuchungsgegenstand waren.

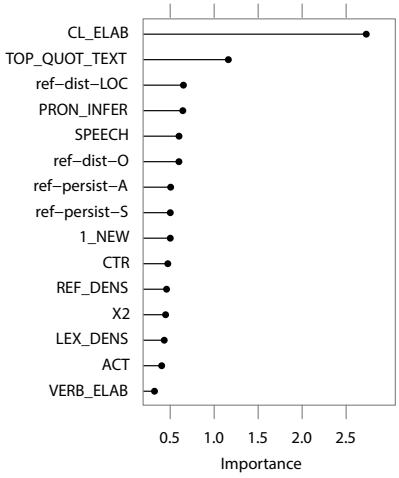
Gesamt-Feature-Sets auch die Differenzierung einer solchen diskurspragmatischen Typisierung ermöglichen können.[59]



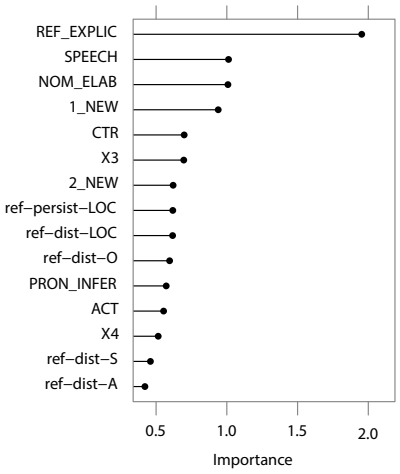
Plot 6.6.10: Feature-Importance Top 15 für BASE-Klassen (Gesamt-Feature-Set)



Plot 6.6.11: Feature-Importance Top 15 für GENRE-Klassen (Gesamt-Feature-Set)



Plot 6.6.12: Feature-Importance Top 15 für COMM_SIT-Klassen (Gesamt-Feature-Set)



Plot 6.6.13: Feature-Importance Top 15 für DISC_STRUCT-Klassen (Gesamt-Feature-Set)

59 Die diskursfunktionale `DISC_STRUCT`-Klassifizierung verläuft z. T. quer zu der Textsorten-Einteilung: So sind z. B. einige dort als Erzählung klassifizierte Texte hier als verhaltensbezogener Diskurs (und nicht als narrativer) eingeordnet, da sie auf eine logische Abfolge bezogen sind – so etwa bei den `Misbehave`-Texten oder den Varianten von `OldDog(Wo)Man` (*Wenn man nachts arbeitet (spinnt), kommt das Ungeheuer* usw.), die zwar durchaus auch längere chronologische Handlungsabfolgen beinhalten, aber eine ermahnend-verhaltenssteuernde Kommunikationsabsicht aufweisen.

6.6.2 Gesamtbewertung der Sequenzanalysen

Neben den Feature-Set-bezogenen textstrukturellen TWM-Parametrisierungen wurden in dieser Arbeit auch Sequenzanalysen erprobt, die als kategoriale bzw. numerische Folgen (insbesondere **Partitur-Folgen** im Sinne von Schulze 2018; 2019; vgl. 3.1.2) Eigenschaften der **globalen sequentiellen Textstruktur** modellieren (im Gegensatz zu den Frequent-Pattern-Daten von häufigen Ereignisübergängen, die *lokale* Sequenzinformationen darstellen und die über ihre Textfrequenz als Teil des Feature-basierten Gesamtmodells ausgewertet wurden).

Als Parameter für ein globales textstrukturelles Sequenz-Alignment wurden dabei analysiert:

- **Ereignistyp-Sequenzen** als kategoriale Verbklasse-*tag*-Sequenzen bzw. -Übergangssequenzen
- **Switch-Reference-Struktur** als kategoriale *tag*-Sequenzen des Switch-Reference-Status von Subjekten
- **Textinterne Diskursstruktur** als kategoriale *tag*-Sequenzen des direkte-Rede-Status von prädikativen Einheiten
- **Textinterne Diskursstruktur** als numerische Partitur-Folgen des binär kodierten bzw. frequenzaggregierten direkte-Rede-Status von prädikativen Einheiten.

Da der Schwerpunkt der textstrukturellen Operationalisierungen in dieser Arbeit auf den merkmalsbezogenen Analysen liegt, deckt die Auswahl an Sequenzanalysen hier die Bereiche der referentiellen, relationalen und informationsstrukturellen kognitiven Textorganisation nur ausschnittsweise ab; da also die vorgestellten sequentiellen Analysen primär der Methodenevaluation dienen, wird hier auf eine Durchführung von multivariaten Sequenzanalysen für eine kombinierte Auswertung der sequentiellen TWM-Merkmale verzichtet.[60] Stattdessen soll eine kurze Bewertung der in den Sequenzanalysen angewendeten Verfahren erfolgen (in Abschnitt 6.7.2 folgt dann eine abschließende Zusammenfassung der Ergebnisse der durchgeführten Sequenzanalysen).

Die vergleichende Analyse der kategorialen sowie der Partitur-Operationalisierung des Parameters eingebetteter Diskursstrukturen in 6.5.6 und 6.5.7 zeigt, dass sich **DTW**-Distanzen grundsätzlich zur Analyse von Textpartitur-Folgen (im Sinne von Schulze 2019 als Sequenz aggregierter Frequenzen einer linguistischen Einheit bzgl. einer höheren Einheit) eignen und hier meso- bzw. makrostrukturelle Sequenzmustertypen feststellbar sind. Die Ergebnisse der Clusteranalysen textstruk-

60 Für eine multivariate Clusteranalyse von kategorialen Sequenzdaten können kategoriale Multichannel-Sequenzanalysen (Gabadinho u. a. 2011) verwendet werden; für Partitur-Folgen können im Rahmen der DTW-Berechnung multivariate Dynamic-Time-Warping-Analysen numerischer Folgen durchgeführt werden (vgl. Giorgino 2009: 12ff.).

tureller Sequenzen als *tag*-Folgen mit **Optimal-Matching**-Distanzmaß und *max-length*-Normierung sind dagegen größtenteils schwierig zu interpretieren; so werden z. B. in der Analyse von Phasen direkter Rede offenbar kleinteiligere lokale Ähnlichkeiten berücksichtigt und nicht die globale textweite Struktur (vgl. 4.4.4; Aggarwal 2015: 503).

Bei der Klassifikation mit **Spectrum-String-SVM-Kernel** für kategoriale Sequenzmodelle für die Vorhersage der Apriori-Textkategorisierungen mit den verschiedenen sequentiellen Modellen (Ereignissequenzen, Switch-Reference und Diskursstrukturen) zeigt nur das Ereignissequenzmodell ein Kappa > 0.4 (*moderate agreement*), und zwar für die (binäre) Klasse `COMM_SIT` (für die knn-Klassifikation der DTW-Modelle wurde kein Kappa bestimmt). Es bleibt zu prüfen, inwiefern sich diese Ergebnisse durch Anwendung multivariater Sequenzklassifikationsmethoden (vgl. Kuksa 2014) auf einem umfangreicheren Set an Sequenzen – analog zur Feature-basierten Klassifikation mit dem Gesamt-Feature-Set – verbessern lassen.

6.7 Diskussion der Ergebnisse

6.7.1 Diskussion der Feature-Analysen

In der Clustertypologie des Gesamt-Feature-Sets der hier als potentielle TWM-Parameter untersuchten Merkmale werden die im Korpus gegebenen **Märchenvarianten** jeweils gemeinsam gruppiert (s. schwarze Markierungen in Abbildung 6.6), d. h. als Typen mit ähnlicher Textstruktur identifiziert.[61] Im Einzelnen werden folgende Varianten gruppiert (Auflistung von rechts nach links im Dendrogramm 6.6):[62]

- `Cranberry` (`anim_tal`): Varianten aus zwei Khanty-Dialekten (Surgut-Khanty, 1992, und Yugan-Khanty, 2012/2015); minimal in der Konzentration auf die zwei Gegenspieler (Cranberry und Grass Bundle), fabelartig (moralisch-belehrend: ‚Schadenfreude ist schlecht'); bilden eigenen Schwestercluster (2) zu der Kerngruppe (1).[63]
- `BeardedMan` (`anim_tal`): Nordmansi-Variante (erste Hälfte 20. Jh.) und Surgut-Khanty-Variante (1992); ähnliches Handlungsstrukturmuster wie bei `Cranberry` (‚Gefahr durch Verbrennen', ‚Alle sterben'), allerdings mehr Aktanten.

61 Nur die beiden Varianten der Yugan-Khanty-Erzählung `Misbehave` sowie einer der `TwoFires`-Texte werden nicht mit ihren entsprechenden Varianten geclustert, sondern mit anderen Märchen desselben Dialekts; außerdem enthält der Cluster mit den beiden `LittleBird`-Märchen auch die persönliche Erzählung `Cold`, was u. a. der deutlich unterschiedlichen Textlänge der beiden `LittleBird`-Varianten geschuldet sein mag.

62 Vgl. auch die Übersichtstabellen 5.2 und 5.3 zu den Sammlungen sowie das Korpusverzeichnis.

63 Siehe 6.7.2 für den Text 1484 `LittleCranberry(AIK)` in englischer Übersetzung sowie als Interlinearversionen in Fußnote 78.

- `TwoFires` (`anim_tal`): Jugan-Khanty-Varianten (2015), verschiedene Sprecher; kurze, moralisch-belehrende Erzählungen mit zwei Feuerstellen als Gegenspieler (Thema ‚gute Pflege des Feuers').
- `LittleBird` (`magic_tal`): Varianten aus zwei Khanty-Dialekten (Yugan-Khanty, 1901, sowie Surgut-Khanty, 1992); Hybrid zwischen Zaubermärchen (Kampf mit Gegenspieler, dem Mank-Waldgeist) und den kurzen fabelartigen Erzählungen (moralisch-belehrend; in einer Version werden die Protagonisten nach dem Sieg über den Mank zur Strafe von einer Eule gefressen); gruppiert in einem Subcluster mit den `TwoFires`- und `BeardedMan`-Varianten (beide `anim_tal`).[64]
- `OldDog(Wo)man` (`magic_tal`): Jugan-Khanty-Varianten (2012/2015), verschiedene Sprecher; Erzählungen von der Überlistung des Ungeheuers *Amp-Chun-Lonep* (Greis(in)-mit-dem-Hunderückgrat); mit den Mansi-Zaubermärchen subgruppiert.

Da sich die verschiedenen Texte einer Märchenvariante jeweils deutlich bzgl. Textlänge, Motivfolge und sprachlicher Kodierung der Ereignisse unterscheiden,[65] kann man davon ausgehen, dass bei diesen Textvarianten eines Märchens (auch bei Varianten im selben Dialekt) nicht der Text als solcher in seiner Gänze memoriert wurde:[66] Stattdessen ist anzunehmen, dass der Märchenerzähler bzw. die Märchenerzählerin das im Langzeitgedächtnis abgespeicherte Text-Modell des Märchens abrief, das zuvor bei seiner bzw. ihrer ursprünglichen Rezeption des Textes als kognitives Modell der im Text kodierten Informationseinheiten im Arbeitsgedächtnis aufgebaut worden war (s. Schwarz-Friesel & Consten 2014: 58; vgl. 1.1.2 und 3.1.1), und dieses gemäß der TWM-Struktur-Regeln für das Märchen-Genre textuell neu kodierte. Dementsprechend sind die Varianten anzusehen als „spontaneous reformulations of

64 Für Details zum `LittleBird`-Märchen s. Csepregi 2005 („Das Vöglein und seine Schwester – Variationen eines obugrischen Märchentyps"); zum Mank-Waldgeist („the enemy of the people – a harmful spirit living in the forest and having super-power", Kerezsi 1995: 184) s. auch Zehetmaier & Fónyad 2020.
65 Vgl. die in den Cluster-Labels kodierte Anzahl an Sätzen als Textlängenmaß: Z. B. hat die `LittleBird`-Variante 1314 eine Länge von 90 Sätzen, die `LittleBird`-Variante 732 dagegen nur eine von 39 Sätzen.
66 Nur einige formelhafte Wendungen wiederholen sich in bestimmten Märchenvarianten. So teilen sich etwa die `LittleBird`-Varianten die Einleitungsformel:

(13) pi:tʲəŋkəli-ɣən=o:pi-sɐ-ɣən βɑł-ɬ-əɣən
little_bird-DU=older_sister-COLL-DU live-PRS-3DU
There lived a little bird and his older sister. (SK, 732: 1)

sowie folgenden formelhaften Satz:

(14) mɛŋk i:ki jɑqqn əntem
Menk Sir at_home NEG.EXIST
The maŋk was not at home. (PA, 1314: 6)

narrative traditions“ (Schulze 2004b: 206; vgl. 5.2.3)[67] – sie sind also nicht Varianten eines *Textes* (in diesem Fall wäre die im Clustering festgestellte Ähnlichkeit bzgl. textstruktureller Parameter trivial), sondern sie sind verschiedene Varianten der textuellen Neukodierung eines tradierten *Text-Modells* – und weisen eben deshalb eine sehr ähnliche schematische Struktur auf.

Neben diesen Varianten einzelner Märchen im Sinne eines (ähnlichen) Motivfolgemusters werden auch die **Mansi-Zaubermärchen** (`magic_tal`) mit unterschiedlicher Motivfolge, aber mit vergleichbarem übergeordnetem Handlungsstrukturmuster (Konflikt mit bösem Gegenspieler, Sieg Gut gegen Böse; vgl. Propp 1972; s. auch 5.2.3)[68] gemeinsam gruppiert (orange Markierung in Abbildung 6.6). Diese Gruppe bildet wiederum mit den als Zaubermärchen ausgezeichneten Jugan-Khanty-Erzählungen einen Cluster innerhalb der Kerngruppe (1), den man als anhand der textstrukturellen Features induktiv rekonstruiertes Zaubermärchen-Subgenre interpretieren kann (vgl. Schulze 2020: 592).[69]

Da die Texte dieses **Zaubermärchen-Subgenres** vom gleichen Handlungsstrukturtyp sind und – repräsentiert über die Gesamtmenge Feature-basierter TWM-Parameter in der vorliegenden Operationalisierung als quantitative Parameter der Text-Modell-Struktur – in der Clustertypologie dieses Gesamt-Feature-Sets zusammen gruppiert werden, kann man daraus schließen, dass diese **TWM-Operationalisierung** tatsächlich solche für die Texte dieses Zaubermärchen-Clusters spezifischen Strukturmerkmale abbildet, deren für dieses Subgenre prototypischen Werte gemäß der Grundthese Weltmodell-basierter Sprachverarbeitung in einem entsprechenden **Text-Weltmodell** (TWM) als kognitivem Strukturmodell abgespeichert wären, das jeweils in der Textproduktion dieses Märchentyps aktiviert wird.

Ähnliches gilt für die Märchenvarianten: Während die Zaubermärchen als TWM-basierte Versprachlichungen von Text-Modellen eines Subgenres im Sinne eines über-

67 Hinweise auf diese „Elastizität“ der Erzählungen (Schulze 2020: 605f.; vgl. auch 1.2.2) geben (vgl. Schulze 2004b: 206) in den obugrischen Texten u. a. Code-Switching (z. B. in Text 1514) und Selbstkorrekturen, wie sie etwa in den `Cranberry`-Texten 1483 und 1484 auftreten. Auch von der inhaltlichen Motivfolge unterscheiden sich diese `Cranberry`-Texte z. T. deutlich: So wird in 1488 das Grasbündel vom Wind verweht, während es in den anderen Varianten in Brand gerät. Während die beiden Vöglein in der Yugan-Khanty-Version 1314 von `LittleBird` am Ende zur Strafe von einer Eule gefressen werden, bleibt ihnen dieses Schicksal in der Surgut-Khanty-Version 732 erspart und sie können von den Reichtümern des getöteten Manks leben.

68 Diese Mansi-Zaubermärchen stellen handlungsstrukturell komplexe, ‚klassische‘ Zaubermärchen im Sinne Propps dar, mit den dafür typischen Motiven wie Familie, Geschwister, Bräutigam, Verwandlung, Rettung, Bestrafung (vgl. Propp 1972; Uther 2011).

69 Zwar ist bei den Mansi-Zaubermärchen, die alle aus dem frühen 20. Jh. stammen, auch die Annahme eines dialektal-arealen (Mansi-spezifischen) oder auch diachronen Subtyps möglich. Dass sie aber mit den Yugan-Khanty-Zaubermärchen aus der aktuellen Feldforschung eine größere Zaubermärchen-Gruppe bilden (in der einzig die moralisch-belehrenden Surgut-Khanty-`LittleBird`-Zaubermärchen fehlen, die eher den Tiermärchen nahestehen), spricht für das Vorliegen eines Zaubermärchen-spezifischen Struktur-Subtyps.

geordneten Handlungsstrukturtyps verstanden werden können, sind die Texte einer Märchenvariante als **Nacherzählungen** eines memorierten Text-Modells (s. 5.2.3) aufzufassen als die TWM-basierte Versprachlichung des gleichen bzw. eines sehr ähnlichen Text-Modells. Entsprechend zeigt ihre im Clustering festgestellte Ähnlichkeit, dass innerhalb des narrativen Genres noch spezifischere Textgruppen als Strukturtypen unterhalb von Subgenres mit sehr ähnlichem Aufbau ihrer kognitiven Text-Modelle identifizierbar sind, dass also die Merkmale der kognitiv-strukturellen (d. h. der globalen, referentiellen, relationalen und informationsstrukturellen) TWM-Parameter auch diese Varianten mit gegenüber Subgenres spezifischerem Wertebereich im Merkmalsraum (vgl. Schulze 2020: 592)[70] als **feinkörnigste Clusterpartitionen** abbilden (vgl. die schwarz markierten Gruppen in Abbildung 6.6, insbesondere den `OldDog(Wo)man`-Cluster als Untergruppe in einem übergeordneten `magic_tal`-Cluster zusammen mit dem orange markierten Cluster der klassischen Zaubermärchen).

Dies deutet darauf hin, dass die als TWM-Parameter angenommenen Merkmale der Referenten-, Handlungs- und Informationsstrukturierung relevante Muster für die strukturelle Kodierung textueller kognitiver Modelle abzubilden vermögen. So zeigen sich im Clustering diese schematischen textstrukturellen TWM-Dimensionen für die Varianten eines Märchens – entsprechend der Annahme, dass diese ein ähnliches, ggf. auch über lange Zeiträume und Sprachgrenzen hinweg tradiertes Text-Modell mit entsprechend damit verknüpftem TWM kodieren – als relativ unabhängig von dem Stil des Sprechers,[71] der Sprache bzw. dem Dialekt, der Zeitstufe sowie auch der Textlänge (s. insbesondere den `BeardedMan`-Cluster in Abbildung 6.6 mit Mansi-Variante aus der ersten Hälfte des 20. Jh.s und Khanty-Variante vom Ende des 20. Jh.s). Damit geben die Auswertungsergebnisse dieser Arbeit also einen datengestützten Hinweis darauf, dass die bei Schulze (2018; 2019; 2020) entwickelten und hier korpusbasiert-explorativ untersuchten Parameter als adäquate **Modellierungen von Text-Weltmodellen** im Sinne von Strukturmodellen der kognitiven Textverarbeitung und -produktion angesehen werden können.

Betrachtet man das Clustering-Resultat der TWM-Operationalisierung durch das Gesamt-Feature-Set in 6.6.1 insgesamt, so bilden diese Varianten bzw. Subgenres narrativer Texte jeweils Subgruppen innerhalb eines Hauptclusters mit narrativer Text-

70 Schulze 2020: 592: „Doch steht zu vermuten, dass diese Optionsräume funktionale und formale, in Teilen sicherlich prototypisch organisierte Wissenskategorien spiegeln, denen mehr oder minder komplexe Cluster von sprachlichen Repräsentationsformen zugeordnet sind. Je kleiner die Optionsräume sind und je konkreter weitere Repräsentationsformen typischerweise auftreten, desto spezifischer wird der Genre-Bereich, der dann als ‚Subgenre' bezeichnet werden kann […]."

71 Vgl. die Verteilung der Sprecherangaben (Kürzel in Klammern am Ende der Titelcodes) bei den Yugan-Khanty-Texten in Abbildung 6.6 (Dialektkürzel YK).

struktur (Cluster 1 und 2 in Abbildung 6.6).[72] Die **mythologischen Erzählungen** verteilen sich über verschiedene Untergruppen dieses Hauptclusters, besitzen also kein im Rahmen der Operationalisierung eindeutig von den Volkserzählungen differenzierbares Textstrukturschema. Sie sind zwar thematisch anders gelagert als die Zauber- und Fabelmärchen, teilen aber strukturell Merkmale entweder mit den längeren Zaubermärchen (z. B. 750 als mythische Sage)[73] oder mit den kürzeren Fabel-Tiermärchen (z. B. der Schöpfungsmythos 741 mit dem Motiv des ‚Hochtauchens' der Welt durch den Haubentaucher, s. Interlinearversion 2 in 5.3.6 für den Text).

Die **ethnographischen Berichte** sind bis auf eine Ausnahme (s. u.) Teil des narrativen Kernclusters (1); das kann man dahingehend interpretieren, dass in der Text-Modell-Konstruktion dieser *nicht-fiktiven* Texte ein ähnliches narratives Erzählschema eingesetzt wird wie für fiktive Erzählungen, sodass man diese Texte auch strukturdatengestützt als **persönliche Erzählungen** typisieren kann. Auch der journalistische Text 1348 (`Guest`) hat als *personenbezogene* chronologische Nacherzählung von Reiseetappen eine narrativen Texten ähnliche bewegungsbezogene Ereignistypik (s. Plot H.3) und wird ebenso dem Kerncluster (1) zugeordnet. Gekennzeichnet ist dieser **narrative Kern-Texttyp** (1 = rot) allgemein u. a. durch eine stark **handlungsbezogene** Ereignistypik (bzw. Verkettungen von Bewegungs- und Handlungsereignissen) und eine **niedrige** relative referentielle **Informationsdichte** (s. Plots 6.7.1f.; vgl. die Auswertung in 6.6.1).[74]

72 Gruppe 1 (= rot) unterscheidet sich von Gruppe 2 (= grün) unter anderem durch das Vorkommen des häufigen Ereignisabfolgemusters `ACTION > MOTION > ACTION`, das in den kurzen `Cranberry`-Erzählungen fehlt; dafür sind diese stärker `ACTION`-orientiert, s. Plot 6.7.2.

73 ATU-Gruppe 750–849 = Legendenartige Märchen bzw. Religious Tales; diese bilden im ATU-Index zusammen mit den Zaubermärchen die Gruppe der Ordinary Tales (Uther 2011).

74 Als Beispiele für Vertreter dieses narrativen Perspektivierungstyps mit Blöcken mit durchlaufendem Thema (s. 3.8.2; Daneš 1970: 76) vgl. zum einen das Beispiel des Zaubermärchens 1237 (`ThreeSons`) in Abschnitt 6.2.2, Fußnote 28; zum anderen den Beginn des nordmansischen Tiermärchens 1346 (`BeardedMan`) – hier wird im zweiten Satz die Passiv-Diathese zur Topikalisierung eingesetzt, um die Hauptreferentengruppe (Referenten-ID `1+2+3`) in Subjektposition zu belassen:

(15)

əj_mətɐ_ɬɐːtnə	tuːʃəŋ_ɐːɣən	pɐːnə	βɐːtʲ_kʉr_βɛjɣən	pɐːn	oːɣ_ɬoɬəŋkən	
some_day	Bearded_Chin	and	Two_Thin_Legs	and	Two_Temples	
[ADV]	[S				]	
-	1+2+3					
βɑɬ-ɬ-ət						
live-PRS-3PL						
[PRED]						
-						

There once lived Bearded Chin and Two Thin Legs and Two Temples. (SK, 1346:1)

(16)

Ø	əj_ɬɐːtnə	ɬʲɐːɬʲ-nə	joβt-ɐt
ø	once	troop-LOC	come+[PST]-PASS.3PL
[S]	[ADV]	[ADV]	[PRED]
1+2+3	-	4	-

One day an army came upon them. (SK, 1346:2)

Abgetrennt von diesem Textstrukturtyp obugrischer Erzählungen, zeigen die **Fate Songs** (`fas`) als „lyric songs, where the moods, experiences and fate of the author (a man or a woman) are described“ (Ojamaa & Ross 2004: 134), ein **andersartiges Strukturschema**, das sich von dem narrativen TWM-Kernschema unterscheidet durch eine hohe Anzahl neuer Topiks, geringe Verweildauer eines Referenten als Subjekt (also **geringe Topic-Continuity**, vgl. Cluster 1 in Plot 6.7.4),[75] eine geringe Frequenz von Handlungsverben und eine hohe Frequenz von Verben der Wahrnehmung (s. Plots 6.7.1f.; vgl. auch Plots H.3ff. sowie die Auswertung in 6.6.1).

Auch der **Zeitungstext** `Faraway` (1231) als *thema-* bzw. *ortsbezogener*, logisch-beschreibender Reisebericht[76] sowie der ethnographische Text `Trap` (1355) – eine

75 Vgl. das Beispiel des Fate Songs 1368 (`GameVoicedSong`) in Abschnitt 6.6.1, Fußnote 51, mit dem Muster eines Subjektwechsels zwischen Wahrnehmendem und Wahrnehmung.

76 Vgl. folgende Sätze dieses mansischen Zeitungstextes (1231 `Faraway`) mit linearer Progression statt durchlaufendem Thema (vgl. 3.8.2, Fußnote 26; s. auch Daneš 1970: 75f.):

(17)

ma:n	lʲudvʲig-maksʲimʲilʲian	nampa	unʲivʲersʲitʲet-ət	mʲunxʲen	u:s-t
1PL	Ludwig-Maximilian	named	[Ru._university]-LOC	[Ru._Munich]	city-LOC
[S]	[ADV		]	[ADV	]
1	3			4	

xanʲisʲtaxt-as-uw
learn-PST-1PL
[PRED]
-

We studied at the Ludwig-Maximilian University in the city of Munich. (NM, 1231:2)

(18)

tot	fʲinnougrovʲedʲenʲie	os	uralʲistʲika	xanʲisʲt-an	kol
there	[Ru._Finnougristics]	and	[Ru._uralistics]	teach-PTCP.PRS	house
[ADV]	[SUBPRED			]	[S]
-	-				5

o:l-i
be-PRS[3SG]
[PRED]
-

There is an institute for Finno-Ugric and Uralic Studies there. (NM, 1231:3)

(19)

ma:n	ru:t	la:tŋ-anuw	-Px	nʲemʲets	ma:xm-ət-n
1PL	kinsfolk	language-PL<1PL	-PX	[Ru._German]	people-PL-DLAT
[S		]	[MOD]	[ADV	]
6			1	7	

pu:msʲal-awe-t #
to_be_interested_in-PASS-3PL
[PRED]
-

Ø	os	xanʲisʲt-awe-t
ø	and	teach-PASS-3PL
[S]	[CON]	[PRED]
6	-	-

Germans are interested in our related languages and teach them. (NM, 1231:4)

Art Wegbeschreibung[77] – werden in der hier gegebenen Operationalisierung als nicht-handlungsbezogene, sondern **expositorisch-informierende** Texte (vgl. Biber 1992b: 231f.) im Clustering diesem im Korpus induktiv festgestellten, **nicht-narrativen Textstrukturtyp** (3 = blau) mit **hoher** referentieller **Informationsdichte** (vgl. Plot R.8) zugerechnet.

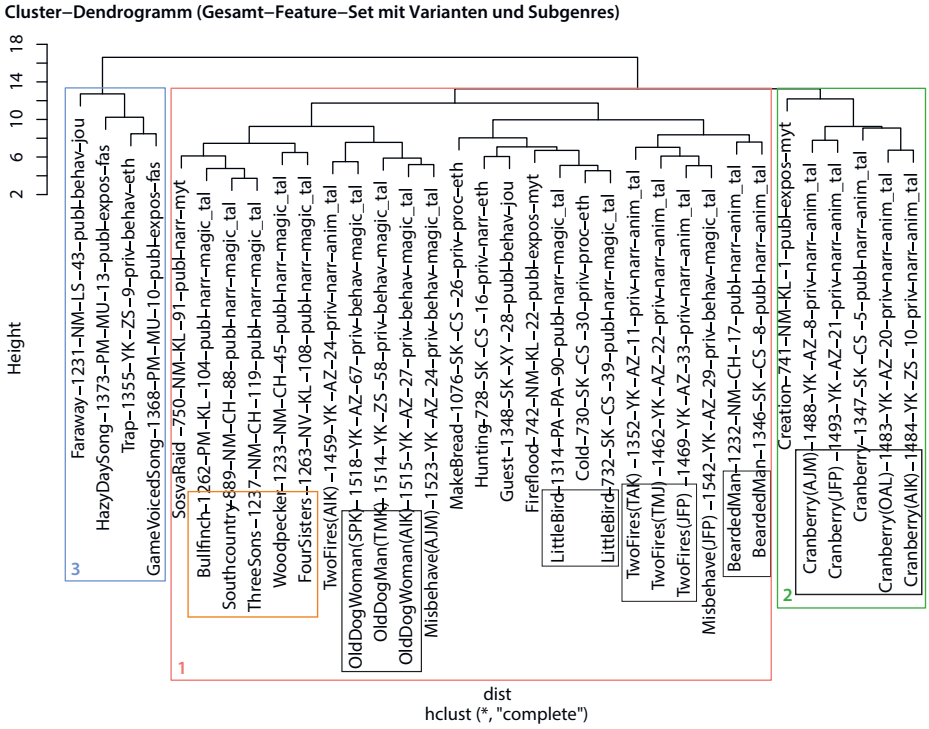


Abbildung 6.6: Varianten (schwarz) und Subgenres (orange) im Cluster-Dendrogramm des Gesamt-Feature-Sets

77 Vgl. die letzten beiden Sätze dieses Yugan-Khanty-Textes (1355 `Trap`):

(20)

tʲuː	jɒːm	tompinə	tot	sɐːp	sɐːp	totti
that	pine_forest	behind	there	tributary	tributary	is_there

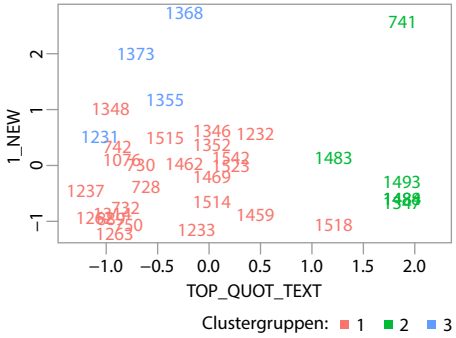
Behind that pine forest is a little tributary... there's a little tributary. (YK, 1355: 8)

(21)

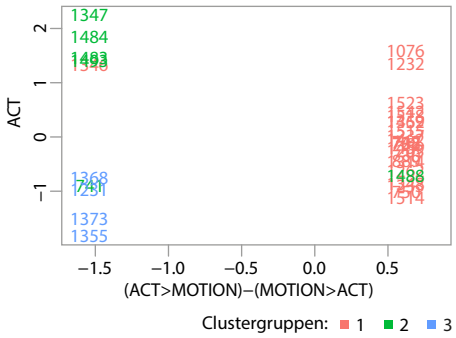
top	sɐːp-ɐ	joɣt-t-ən	məttə	ənəɬ	ontʃəɣ
as_soon_as	tributary-DLAT	come-PRS-2SG	(s)he_says	big	pine_tree

porəq-nɐt	otə	βɛɬtip	ɒːməs-ɬ
bulky_end_of_a_tree_next_to_the_roots-COM	ehr	trap	sit-PRS[3SG]

When you come to the little tributary – he says – under the roots of the large pine tree, ehr, there's the trap. (YK, 1355: 9)



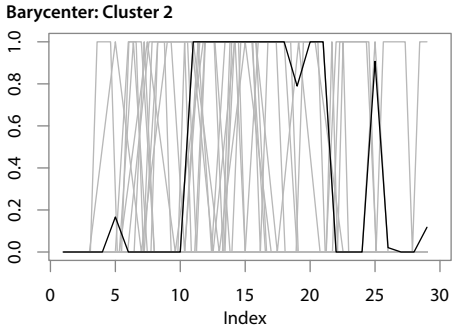
Plot 6.7.1: Topikalitätsmerkmale nach Clustergruppen (Gesamt-Feature-Set)



Plot 6.7.2: Handlungsbezogene Features nach Clustergruppen (Gesamt-Feature-Set)

6.7.2 Diskussion der Sequenzanalysen

In der Auswertung der Ergebnisse der sequenzbasierten Modellierungen von TWM-Strukturparametern zeigt sich, dass insbesondere durch **Dynamic-Time-Warping**-Analysen numerischer Partitur-Folgen makro- bzw. mesostrukturelle narrative Muster identifizierbar sind.



Plot 6.7.3: Barycenter des Clusters mit actio-reactio-Mittelteil (Binäre DTW-Diskurspartitur)

So kann im DTW-basierten Clustering (s. 6.5.7) etwa ein bei den meisten fabelartigen Tiermärchen (vgl. 5.2.3) auftretender **Dialog-Mittelteil** aus Rede und Gegenrede identifiziert werden, der hier als der actio-reactio-Mittelteil des vierteiligen Aufbauschemas der Fabel interpretiert wird (vgl. Dithmar 1974: 104f.). Plot 6.7.3 zeigt die DTW-Kurven des sequentiellen Verlaufs von Dialogstrukturen für die Texte dieser Clustergruppe (Gruppe 2 in Plot Q.1), zusammen mit einem Barycenter-Durchschnittsplot (schwarze Kurve; vgl. 6.1.5).

Ein im Clustering als Vertreter dieses actio-reactio-Sequenzverlauftyps identifizierter Text ist das kurze Tiermärchen 1484 `LittleCranberry(AIK)`, vgl. die englische Übersetzung mit hervorgehobenem Dialog-Mittelteil:

Little Cranberry and Grass Bundle live.
*And Grass Bundle says: **"I, well..."**, well, Cranberry, something like this, however, Little Cranberry says: **"I, well..."***
*Grass Bundle says: **"I'm freezing."***
*Little Cranberry says: **"Hey, light the fire up!"***
Little Cranberry and..., Grass Bundle lit up the fire.
And he was ignited by the fire.
Just [in the moment] when he lit up the fire.
And just when Little Cranberry began to laugh, he starved to death.
They burst.
The End.[78]

Anders als die DTW-Analysen, in denen als numerische Sequenzen repräsentierte Texte gestaucht (Warping) und so aufeinander abgebildet werden, zeigen sich die durchgeführten **kategorialen Sequenzanalysen** in der explorativen Datenanalyse mit OM-Distanzmaß und *maxlength*-Normierung als nur bedingt geeignet, solche globalen textstrukturellen Muster in Texten unterschiedlicher Länge zu identifizieren

78 Ergänzend Text 1484 `LittleCranberry(AIK)` im Original:

(22) pɐːn_sɛməli pɐːn poːm_muːntəl βoɬ-t-əɣən
little_cranberry and grass_bundle live-PRS-3DU

(23) pɐːnə poːmi_muːntəl jɐːstə-t mɐː tʲɑqɐ ɐː pɐːn mətɐ pi
and grass_bundle say-PRS[3SG] 1SG well well cranberry some EMPH
tʲit ɑntɐ pɐːn_sɛməli jɐːstə-t mɐː tʲɑqɐ
this however little_cranberry say-PRS[3SG] 1SG well

(24) poːm_muːntəli jɐːstə-t mɐː poːt-tɐː βɛr-ojəm
grass_bundle say-PRS[3SG] 1SG freeze-INF do+[PST]-PASS.1SG

(25) pɐːn_sɛməli-nə βɛr-t-i nɐːj ʉɬ-ɐ kəʃ
little_cranberry-LOC say-PRS-PASS.3SG fire light_up-IMP.2SG hey!

(26) pɐːn_sɛməli pɐːnə poːm_muːntəl pɐːn nɐːj ʉɬ
little_cranberry and grass_bundle and fire light_up+[PST.3SG]

(27) pɐːn nɐːj-nə βitʲimt-i
and fire-LOC ignite+[PST]-PASS.3SG

(28) kʉʃ nɐːj ʉɬ
just_when fire light_up+[PST.3SG]

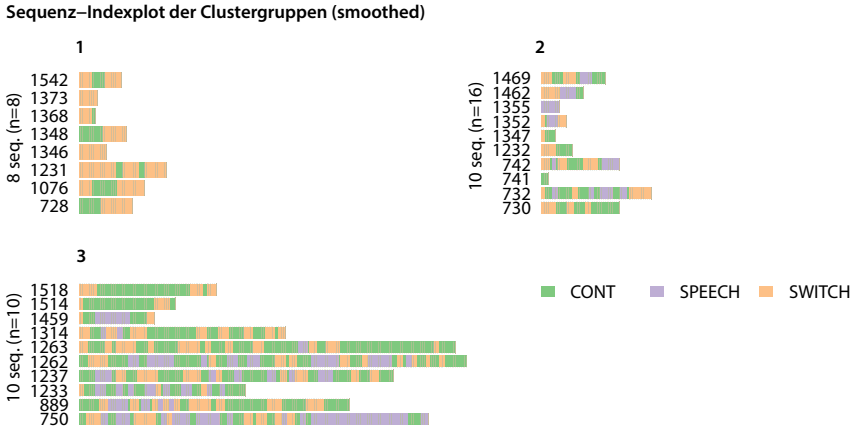
(29) pɐːn_sɛməli pɐːn kʉʃ nʲɑqqətəɣ pɐːn ɐːrɣə
little_cranberry and just_when begin_to_laugh+[PST.3SG] and apart
pɒːjəmtəɣ
starve+[PST.3SG]

(30) əjnɐm pi iːt= puːqən-ɣən
all EMPH PFV= burst+[PST]-3DU

(31) tərəm
end+[PST.3SG]

(vgl. 6.6.2; vgl. auch 4.4.4; Aggarwal 2015: 503).[79] Insbesondere als Mittel der deskriptiven Datenanalyse haben sich solche kategorialen Textstruktur-Repräsentationen jedoch für die Exploration sequentieller Textstruktureigenschaften in dieser Arbeit als äußerst hilfreich erwiesen, insbesondere zur gruppierten Darstellung sequentieller Informationen.[80]

Ein Ansatz zur Verbesserung der Performance von kategorialen Sequenzrepräsentationen im Rahmen der Induktion von TWM-Sequenz-Mustern liegt in einer **Glättung** (Smoothing) zur Reduktion der Zustandswechsel. Für Switch-Reference-Folgen wurde diese durch numerische Kodierung (`SWITCH`: 1, `CONT`: 0, `SPEECH`: -1) und anschließende Anwendung einer Glättungsfunktion (vgl. 4.4.1.2) sowie darauffolgende Rekodierung in die kategorialen Ausgangskategorien erreicht.



Plot 6.7.4: Sequenz-Indexplot für Clustergruppen (Switch-Reference-Sequenzen mit Smoothing)

So trennt diese Operationalisierung (s. Plot 6.7.4) der textweiten Switch-Reference-Struktur, die auch Dialog-Passagen mitberücksichtigt, über entsprechend geglättete Folgen – sehr ähnlich wie die DTW-Analysen der binären Folgen direkter Rede, s. Plot Q.1 – im **Clustering (Plot S.1)** eine Gruppe (1) primär nicht-narrativer Genres ohne direkte Rede mit nur kurzen kontinuierlichen Phasen (`jou`, `fas`, einige `eth`) ab von einer Gruppe (2) um die Tiermärchen mit direkter Rede und höherem Anteil kontinuierlicher Phasen (insbesondere gegen Ende der Texte) sowie von

79 So werden in der entsprechenden Clusteranalyse der Dialogstruktur (s. Plot P.1) die kategorialen Sequenzen offensichtlich eher aufgrund ähnlicher Länge gruppiert als aufgrund einer ähnlichen sequentiellen Anordnung von Dialogpassagen; vgl. etwa die beiden `jou`-Texte ohne Dialogteil in Plot P.3, die (getrennt voneinander) zusammen mit Texten jeweils ähnlicher Länge, aber unterschiedlicher Dialogstrukturierung geclustert werden (s. Plot P.2).

80 Z. B. macht der Indexplot 6.5.4 den hohen Anteil von langen Switch-Reference-Phasen bei den `jou`- und `fas`-Texten sichtbar.

der Gruppe (3) der Zaubermärchen (`magic_tal`), die sich durch viele, oft längere kontinuierliche Phasen auszeichnen, häufig im Wechsel mit Dialogphasen.

Im Gegensatz dazu identifiziert die entsprechende Sequenzanalyse nicht-geglätteter Switch-Reference-Sequenzen (s. Plot L.2) hier zwar auch die längeren Zaubermärchen sowie eine Teilgruppe mit Rede-Gegenrede-Mittelteil (Cluster 3 in Plot L.1); eine Differenzierung der nicht-narrativen Texte von den narrativen ist hier allerdings nicht gegeben.

Alternativ zu diesen Modellierungen der globalen sequentiellen Textstruktur über kategoriale oder numerisch kodierte Zustandsfolgen wurden in dieser Arbeit auch verschiedene aggregierende, Auftrittsfrequenzen berücksichtigende Modellierungen sequentieller Informationen untersucht, die sich insbesondere für eine Operationalisierung von TWM als quantitative kognitive Strukturmodelle anbieten: Neben **aggregierten Partitur-Folgen**, die die globale Abfolge des Auftretens von Informationseinheiten im Text als Folge ihrer Frequenzen in übergeordneten textuellen Einheiten modellieren (s. 6.5.7 für eine entsprechende Operationalisierung der Diskursstrukturierung), gehören dazu Feature-Sets **regional differenzierter** Frequenzmerkmale (wie in 6.5.1 für die Stärke von Topik-Einführungen in erster und zweiter Texthälfte als Modellierung des Informationsflusses im Sinne einer Veränderung der Informationsdichte im Verlauf eines Textes) sowie **Frequent-Patterns** als typische lokale Sequenzmuster (so etwa häufige Ereignisabfolgen in 6.4.2). Im Rahmen der Extraktion solcher Frequent-Patterns können diese über ihre Textfrequenzen als quantitative Textmerkmale mit nicht-sequentiellen Merkmalen in einem Feature-Set kombiniert und diese dann gemeinsam hinsichtlich ihrer Eignung als TWM-Parameter evaluiert werden; so erweist sich etwa in der Analyse für das Gesamt-Feature-Set in 6.6.1 (vgl. Plot 6.6.3) das Ereignisabfolgemuster `ACTION > MOTION > ACTION` als ein zentrales Merkmal zur Differenzierung der im Korpus identifizierten Clustergruppen.

7 Fazit

Forschungsfrage. Ziel dieser Arbeit war die Erprobung von automatischen Verfahren der Mustererkennung und der explorativen Feature-Analyse zur Evaluation von wissensbezogenen textstrukturellen Parametern als Operationalisierung von **Text-Weltmodellen** (**TWM**) im Sinne von genrespezifischen, durch Typisierung von Sprachgebrauchssituationen erlernten, schematischen Strukturmodellen der menschlichen Kognition (vgl. Schulze 2018; 2019; 2020). Neben einfachen quantitativen textstatistischen Maßen mit kognitiver Interpretation, wie der Clauselänge als Elaborationsmaß, der referentiellen Dichte oder der referentiellen bzw. relationalen Explizitheit, wurden als Parameter einer solchen quantitativen kognitiven Texttypologie vor allem referenz- und relationssemantische sowie informationsstrukturelle Merkmale des Aufbaus textuell kodierter kognitiver Modelle untersucht; dazu gehören u. a. die referentielle Distanz, der Topikalitätsquotient, die Ereignistypik, häufige Ereignisschemata, Informationsfluss und -dichte, die textuelle Perspektivierungsstruktur und Muster textinterner Diskursstrukturierung.

Vorgehen. Dazu wurde ein zeitlich und dialektal geschichtetes, syntaktisch, semantisch und informationsstrukturell annotiertes Korpus von 34 Texten obugrischer Volkserzählungen sowie weiterer, hauptsächlich mündlicher Genres (etwa mythologische Sagen, persönliche Erzählungen oder journalistische Berichte) mit automatischen **Klassifizierungsmethoden** des maschinellen Lernens ausgewertet. Unter Anwendung von Methoden der Feature-Extraction sowie des Data-Minings wurden zunächst aus den Annotationsdaten multivariate Feature-Sets sowie Folgen sequentieller Muster als quantitative Operationalisierungen der kognitiven Texttypologie-Parameter extrahiert. Auf diese strukturellen Textrepräsentationen wurden anschließend hierarchische Clusteranalysen für eine induktive Entdeckung kognitiver Strukturmustertypen angewendet. Für Feature-basierte Repräsentationen wurde zusätzlich eine Random-Forest-basierte Feature-Selection für eine Gewichtung dieser Merkmale bzgl. ihrer Relevanz als **TWM-Parameter** durchgeführt.

Ergebnisse. Angelegt als Pretest-Studie zur Methoden- und Feature-Exploration, deutet im Clustering der über die extrahierten textstrukturellen Merkmale repräsentierten obugrischen Texte die gemeinsame Gruppierung der Texte eines narrativen Subgenres von Zaubermärchen sowie von Märchenvarianten[1] über Zeit- und Dialekt-

1 Die Märchenvarianten im Korpus, die textuell sowie in Länge und Motivfolge divergent sind, können als Reproduktion (Nacherzählung) eines im Gedächtnis abgespeicherten kognitiven Text-Modells verstanden werden. Die gemeinsame Gruppierung dieser über textstrukturelle Merkmale repräsentierten Varianten als feinkörnigste Clusterpartitionen innerhalb der Clustergruppen narrativer Texte weist auf eine sehr

grenzen hinweg – zusammen mit der Abtrennung nicht-narrativer Texte[2] wie den Personal Songs – darauf hin, dass die Anwendung der in Kapitel 4 vorgestellten Klassifizierungsmethoden auf die in Kapitel 3 diskutierten Feature-basierten Parameter einer quantitativen kognitiven Texttypologie tatsächlich eine solche Operationalisierung von TWM als **kognitive Genre-Modelle** leisten kann (vgl. Abbildung 6.6).

Die Anwendung von Feature-Selection-Verfahren auf die Clustergruppen zeigt (vgl. Plot 6.6.3), dass für die induktiv im Korpus festgestellten TWM-Strukturtypen insbesondere Merkmale der **referentiellen Informationsstruktur** distinktiv sind, etwa Informationsfluss und -dichte (vgl. Plot 6.6.4) sowie die Topikalitätsverteilung der häufigsten Referenten (vgl. Plot R.2). Daneben sind auch die globalen Textstatistik-Parameter der lexikalischen Dichte sowie der relationalen Explizitheit relevant, ebenso wie die relationsstrukturellen Merkmale der Ereignistypik sowie **lokaler Ereignisabfolgemuster** (so wird im Clustering eine Subgruppe handlungsorientierter Narrationen von solchen mit regelmäßigem Motion-Action-Wechsel als Handlungsstrukturmuster abgetrennt, vgl. Plot 6.6.5).

Die vergleichende Auswertung der Feature-Importance-Daten für verschiedene theoriegestützte Genre-Kategorisierungen zeigt für zwei **Textsorten**-Einteilungen sowie eine binäre Differenzierung (narrativ vs. nicht-narrativ) jeweils topikalitätsbezogene Merkmale als wichtigste Features; darin ähneln sie der induktiv festgestellten Clustertypologie, die aber als einzige Gruppierung eine hohe Relevanz der Ereignisabfolge `ACTION > MOTION > ACTION` für ihre Klassendifferenzierung aufweist. Für zwei vergleichend herangezogene diskursbezogene Klassifizierungen zeigen sich dagegen jeweils andere Feature-Schwerpunkte; entsprechend weisen die Clustergruppen auch eine größere Übereinstimmung mit den Textsorten-Einteilungen als mit diesen Klassifizierungen auf: So ist die Clause-Elaboration wichtigstes differenzierendes Feature bzgl. der binären Registereinteilung privat vs. öffentlich und die referentielle Explizitheit wichtigstes Merkmal für eine diskursfunktionale Kategorisierung (vgl. Plots 6.6.10ff.).

In der Auswertung sequentieller Textstruktur-Repräsentationen konnten u. a. **globale textuelle Abfolgemuster** wie die genretypische Dialogstruktur fabelartiger Tiermärchen mit Rede-Gegenrede-Mittelteil (vgl. Plot 6.7.3) über eine Modellierung als numerische Partitur-Folgen mit der Sequenz-Alignment-Methode des Dynamic-Time-Warpings als quantitative, makrostrukturelle TWM-Muster identifiziert wer-

ähnliche übergeordnete Referenten-, Handlungs- und Informationsstrukturierung dieser Varianten hin und spricht für die Annahme eines schematischen Strukturmodells (TWM) als Grundlage ihrer Textproduktion.

2 Ethnographische Texte werden in der TWM-Parameter-basierten Clusteranalyse nicht klar von den narrativen Kerntexten der Volkserzählungen getrennt; dies könnte ein Hinweis darauf sein, dass die Sprecher bei der Elizitierung dieser persönlichen Berichte im Rahmen der Feldforschung in Ermangelung eines entsprechenden Text-Weltmodells für diesen Kommunikationssituationstyp auf das vorhandene Märchen-Weltmodell zurückgreifen – also in Form einer Erzählung berichten.

den; ebenso verschiedene Typen der thematischen Progression über eine Modellierung der globalen textuellen Perspektivierungsstruktur als Switch-Reference-Folgen mit kategorialen Sequenzanalysen (vgl. Plot 6.7.4).

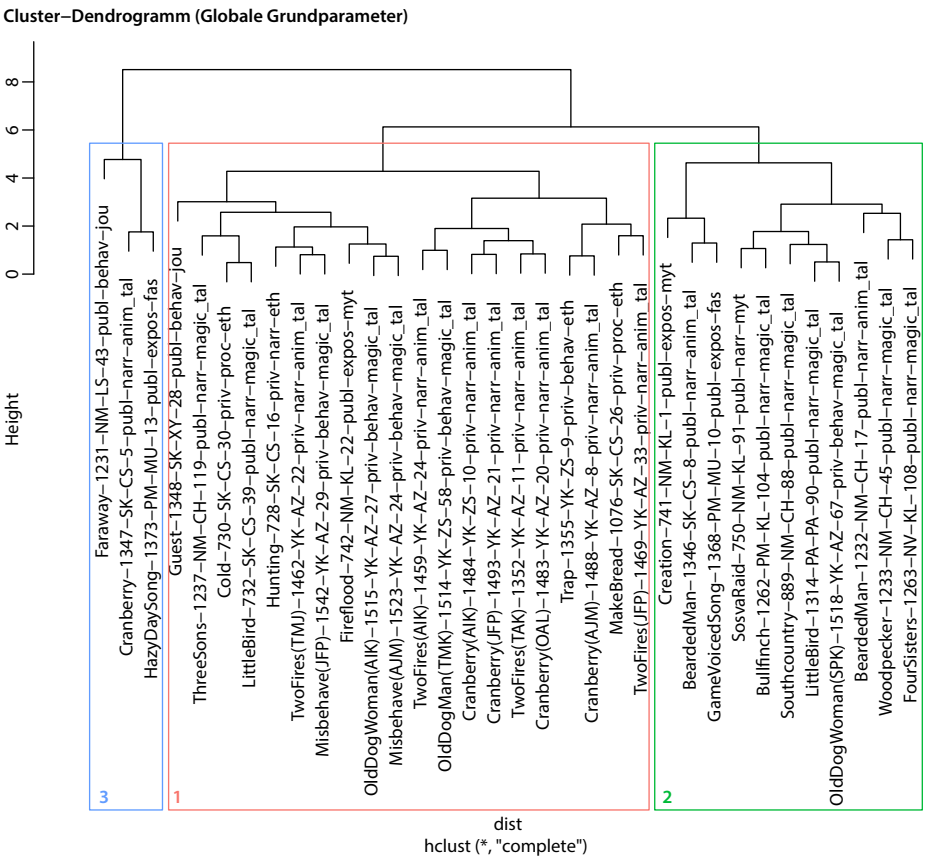
Ausblick. An diese Pretest-Studie, in der sich die grundsätzliche Eignung der getesteten TWM-Operationalisierung und der entsprechenden Extraktions- und Klassifizierungsverfahren zur Identifizierung von TWM-basierten Genre-Mustern gezeigt hat, schließen sich einerseits direkt hypothesenprüfende, kognitiv-linguistische Fragestellungen an, etwa eine vergleichende inferenzstatistische Korpusuntersuchung zum Einfluss der Erzählstruktur russischer Zaubermärchen auf die Textproduktion obugrischer Volkserzählungen. Andererseits können im Anschluss an die in dieser Arbeit anhand eines Erprobungskorpus durchgeführte Methodenexploration die hier getesteten Verfahren einer **TWM-Mustererkennung** auch im Rahmen soziokognitiv orientierter Explorationsstudien (vgl. Schulze 2019: 30f.) auf größere Korpora Anwendung finden, um eine automatische Induktion von Genres als sprachgebrauchsbasierten strukturellen Texttypen zu erreichen. Dabei könnten dann u. a. auch weitere Partitur-Folgen berücksichtigt werden (z. B. die sequentielle Verteilung der automatisch über ihren Topikalitätsquotienten bestimmten Hauptreferenten), die dann mit multivariaten Sequenzanalysemethoden als ‚Stimmen' (Schulze 2019: 15) der TWM-gesteuerten textstrukturellen Komposition, d. h. als die Gesamt-Partitur (Schulze 2020: 627, 629f.) des Textes analysierbar wären.

Auf diese Art sind dann letztendlich verschiedene Text-Weltmodelle als von den Sprechern einer Textkultur geteilte, kognitive Strukturmodelle rekonstruierbar, also als kulturspezifische Diskursordnungen (Foucault 1981: 70f.; 1991: 17ff.), vgl. Schulze 2020: 601: „Diese [*immanenten strukturellen*] Eigenschaften, die auch als Sprachhandlungsroutinen oder -normen verstanden werden können, sind letztendlich Teil der diskursiven Formationsregeln nach Foucault […]." Dies gilt insbesondere für die tradierten Erzählungen einer Sprechergemeinschaft (vgl. Foucault 1991: 18), die gekennzeichnet sind durch eine ihnen eigene Raum-Zeit-Struktur, Referentenstruktur, Handlungsstruktur und Informationsstruktur. In diesem Sinne bleibt mit Lüthi zu schließen:

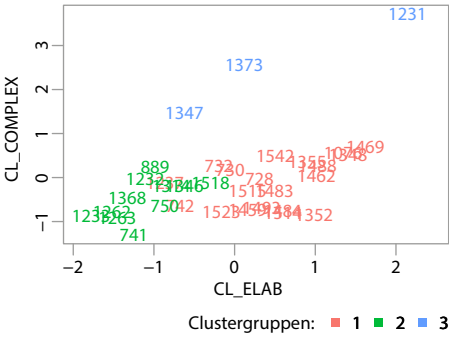
> „Nicht etwa nur von Zeit und Ort", sagt Herder, „binden uns wahre Märchen los, sondern von der Sterblichkeit selbst. Wir sind durch sie im Reiche der Geister." Im Reich der Geister aber herrschen Freiheit und Bindung zugleich. Das Märchen befreit von der Herrschaft und dem Druck äußerer Wirklichkeit, baut aber selber eine wohlgefügte Welt auf, in die es den Hörer oder Leser hereinnimmt. […] Das Märchen besitzt die Freiheit des Gedankens und der Phantasie – aber beide sind nicht gesetzlos, sie fügen sich ganz bestimmten Ordnungen. (Lüthi 1998: 10)

Anhang

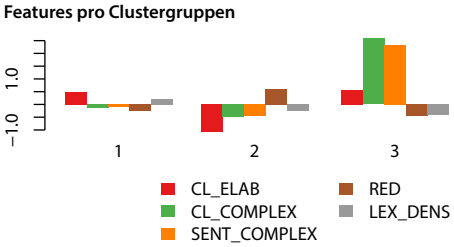
A Plots zu 6.2.1 (Globale Grundparameter)



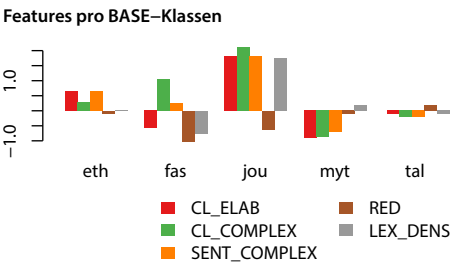
Plot A.1: Cluster-Dendrogramm (Globale Grundparameter)



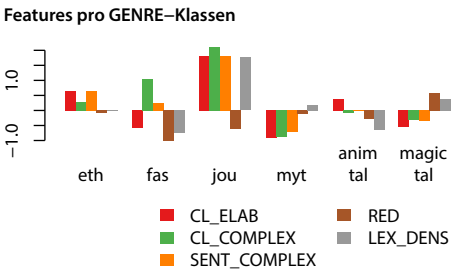
Plot A.2: Clause-Elaboration und -Komplexität nach Clustergruppen (Globale Grundparameter)



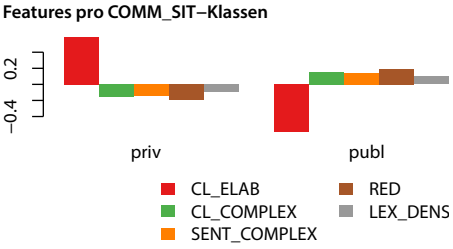
Plot A.3: Clustergruppierte Average-Scores-Barplots (Globale Grundparameter)



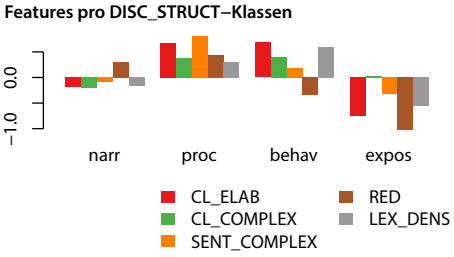
Plot A.4: BASE-gruppierte Average-Scores-Barplots (Globale Grundparameter)



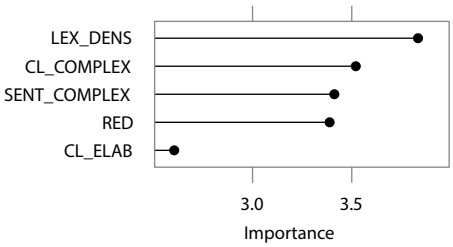
Plot A.5: GENRE-gruppierte Average-Scores-Barplots (Globale Grundparameter)



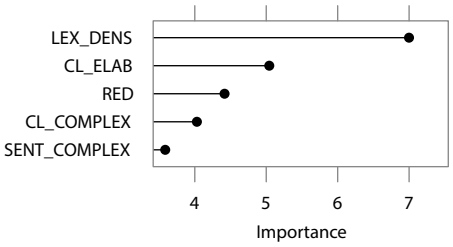
Plot A.6: COMM_SIT-gruppierte Average-Scores-Barplots (Globale Grundparameter)



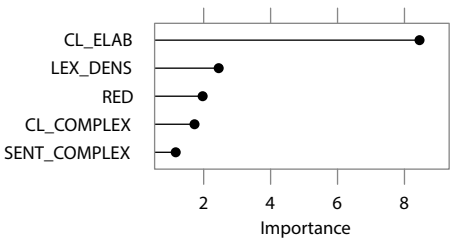
Plot A.7: DISC_STRUCT-gruppierte Average-Scores-Barplots (Globale Grundparameter)



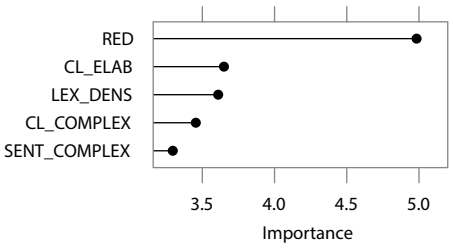
Plot A.8: Feature-Importance für BASE-Klassen (Globale Grundparameter)



Plot A.9: Feature-Importance für GENRE-Klassen (Globale Grundparameter)

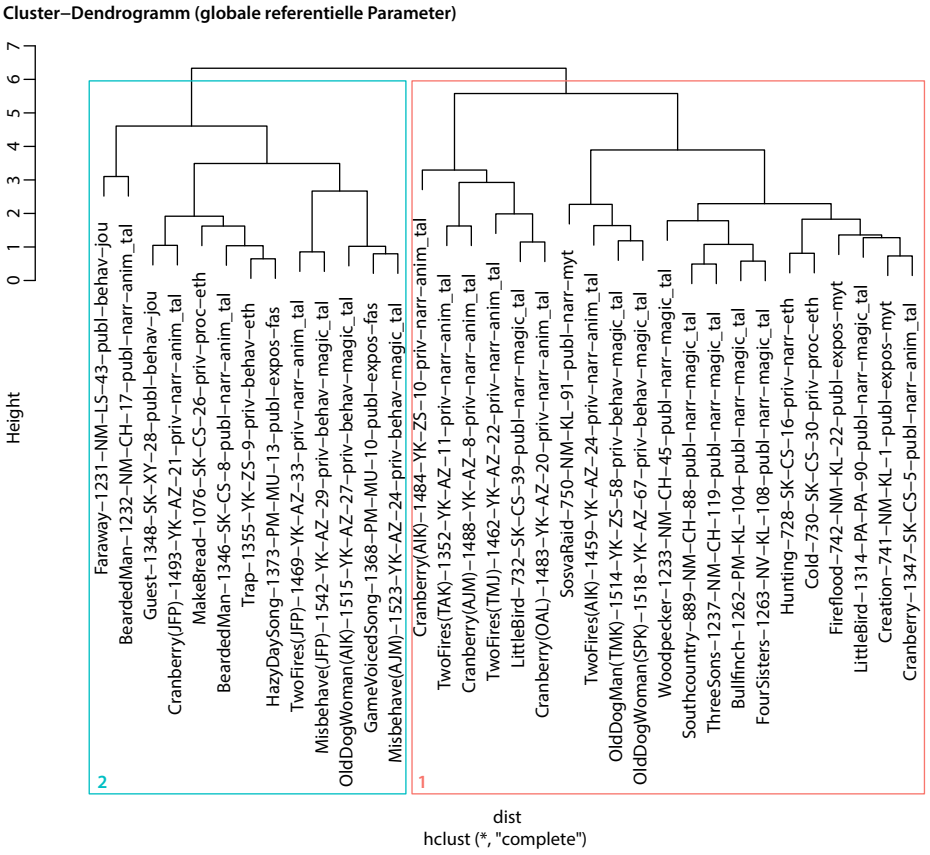


Plot A.10: Feature-Importance für COMM_SIT-Klassen (Globale Grundparameter)

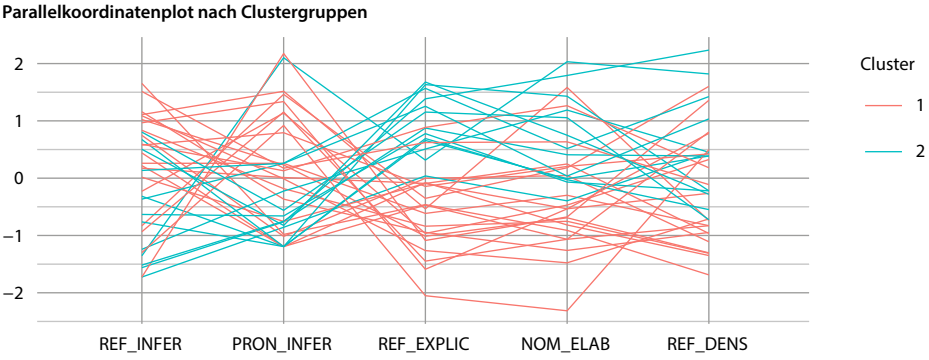


Plot A.11: Feature-Importance für DISC_STRUCT-Klassen (Globale Grundparameter)

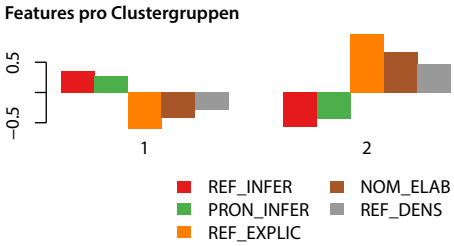
B Plots zu 6.2.2 (Global-referentielle Parameter)



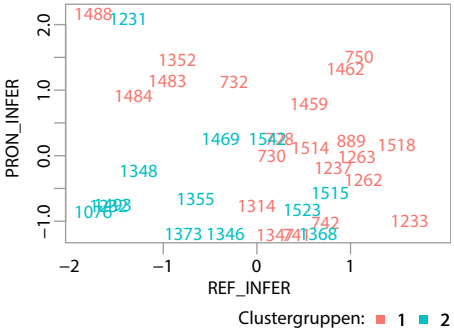
Plot B.1: Cluster-Dendrogramm (Global-referentielle Parameter)



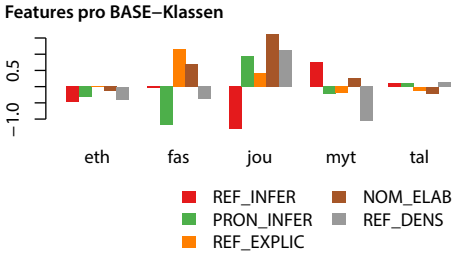
Plot B.2: Parallelkoordinatenplot nach Clustergruppen (Global-referentielle Parameter)



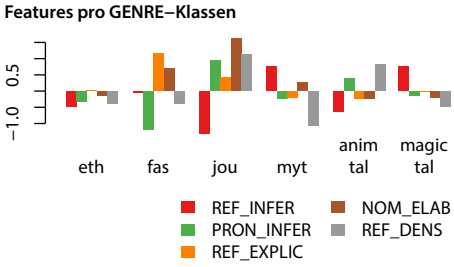
Plot B.3: Clustergruppierte Average-Scores-Barplots (Global-referentielle Parameter)



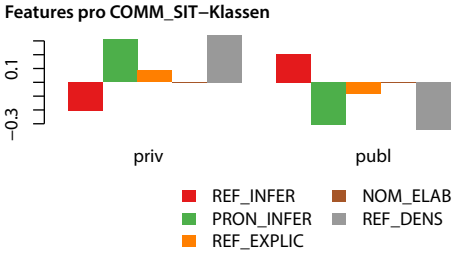
Plot B.4: Referentielle und pronominale Inferenz nach Clustergruppen (Global-referentielle Parameter)



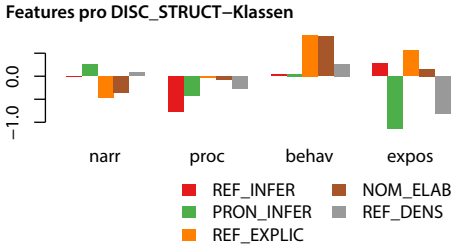
Plot B.5: BASE-gruppierte Average-Scores-Barplots (Global-referentielle Parameter)



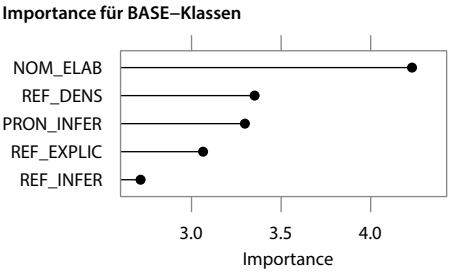
Plot B.6: GENRE-gruppierte Average-Scores-Barplots (Global-referentielle Parameter)



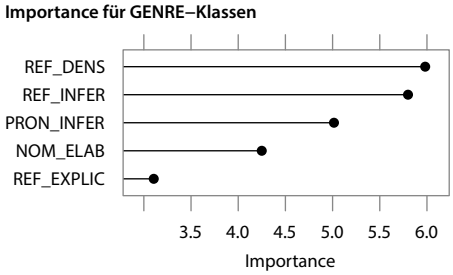
Plot B.7: COMM_SIT-gruppierte Average-Scores-Barplots (Global-referentielle Parameter)



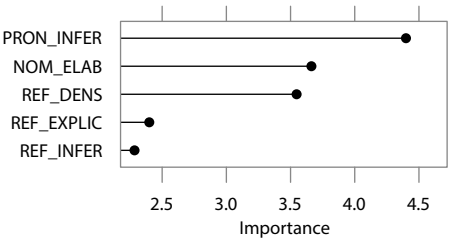
Plot B.8: DISC_STRUCT-gruppierte Average-Scores-Barplots (Global-referentielle Parameter)



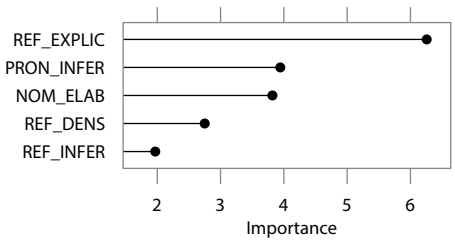
Plot B.9: Feature-Importance für BASE-Klassen (Global-referentielle Parameter)



Plot B.10: Feature-Importance für GENRE-Klassen (Global-referentielle Parameter)

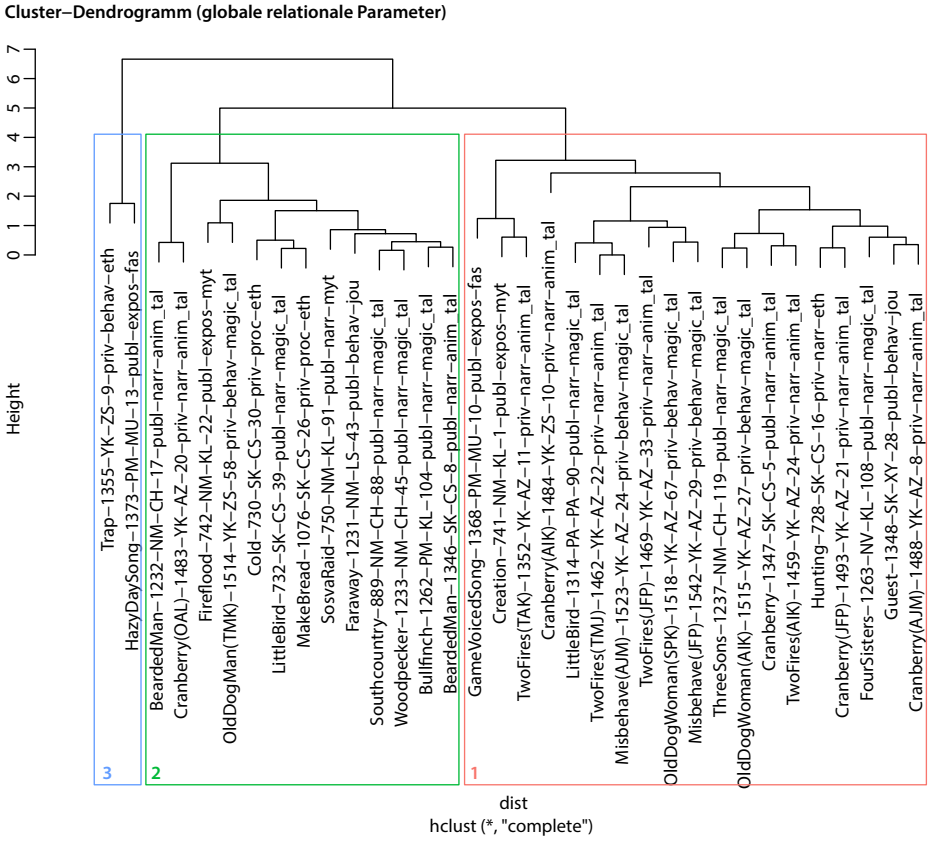


Plot B.11: Feature-Importance für COMM_SIT-Klassen (Global-referentielle Parameter)

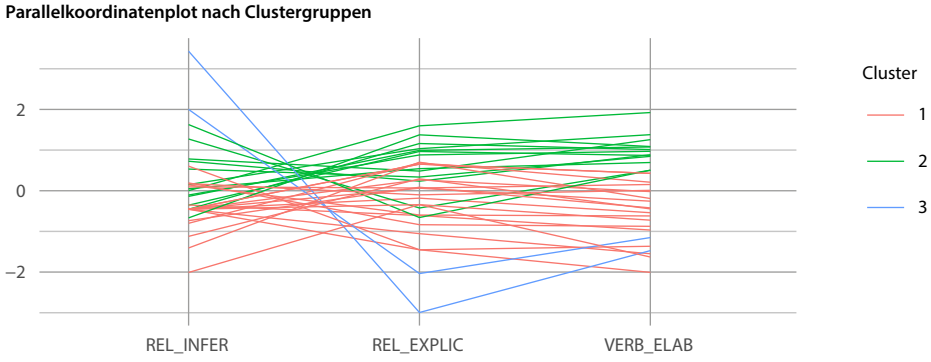


Plot B.12: Feature-Importance für DISC_STRUCT-Klassen (Global-referentielle Parameter)

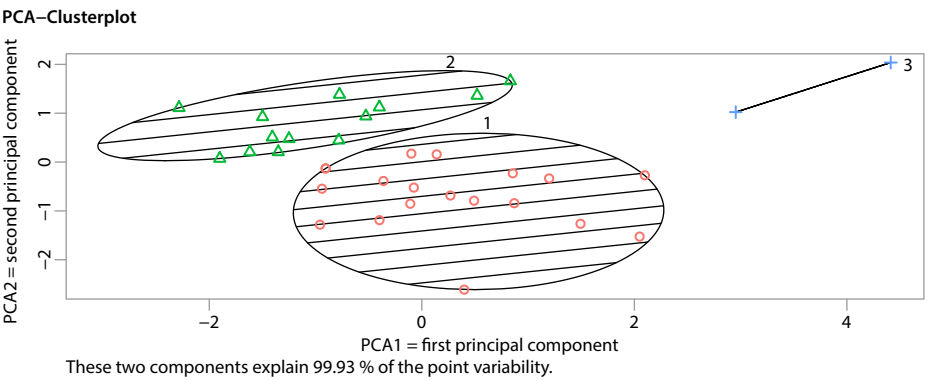
C Plots zu 6.2.3 (Global-relationale Parameter)



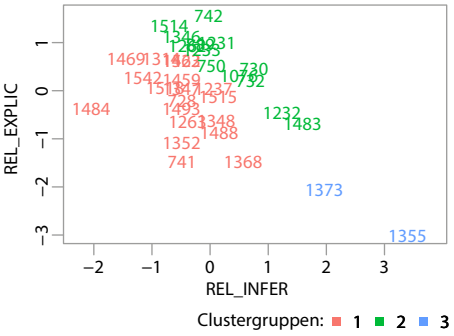
Plot C.1: Cluster-Dendrogramm (Global-relationale Parameter)



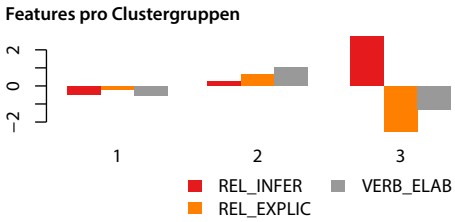
Plot C.2: Parallelkoordinatenplot nach Clustergruppen (Global-relationale Parameter)



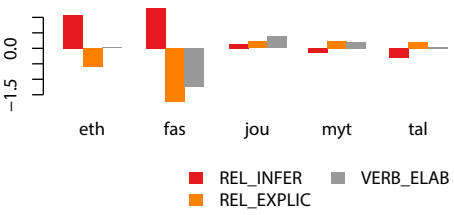
Plot C.3: Hauptkomponenten-Clusterplot (Global-relationale Parameter)



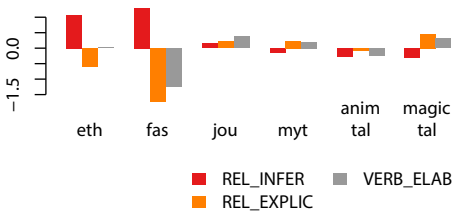
Plot C.4: Relationale Inferenz und Explizitheit nach Clustergruppen (Global-relationale Parameter)



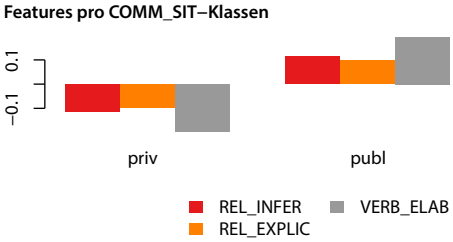
Plot C.5: Clustergruppierte Average-Scores-Barplots (Global-relationale Parameter)



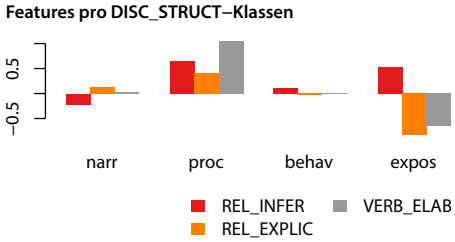
Plot C.6: BASE-gruppierte Average-Scores-Barplots (Global-relationale Parameter)



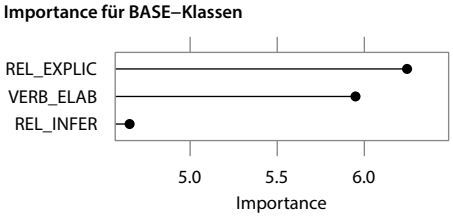
Plot C.7: GENRE-gruppierte Average-Scores-Barplots (Global-relationale Parameter)



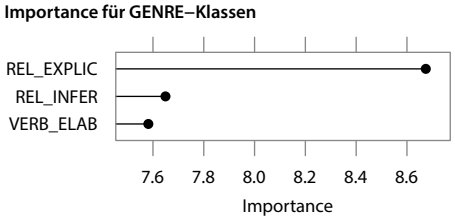
Plot C.8: COMM_SIT-gruppierte Average-Scores-Barplots (Global-relationale Parameter)



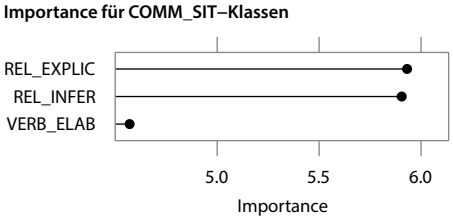
Plot C.9: DISC_STRUCT-gruppierte Average-Scores-Barplots (Global-relationale Parameter)



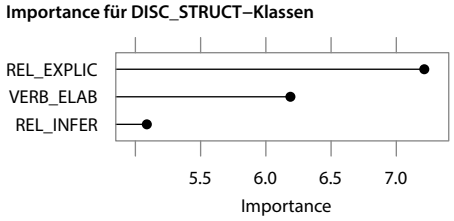
Plot C.10: Feature-Importance für BASE-Klassen (Global-relationale Parameter)



Plot C.11: Feature-Importance für GENRE-Klassen (Global-relationale Parameter)

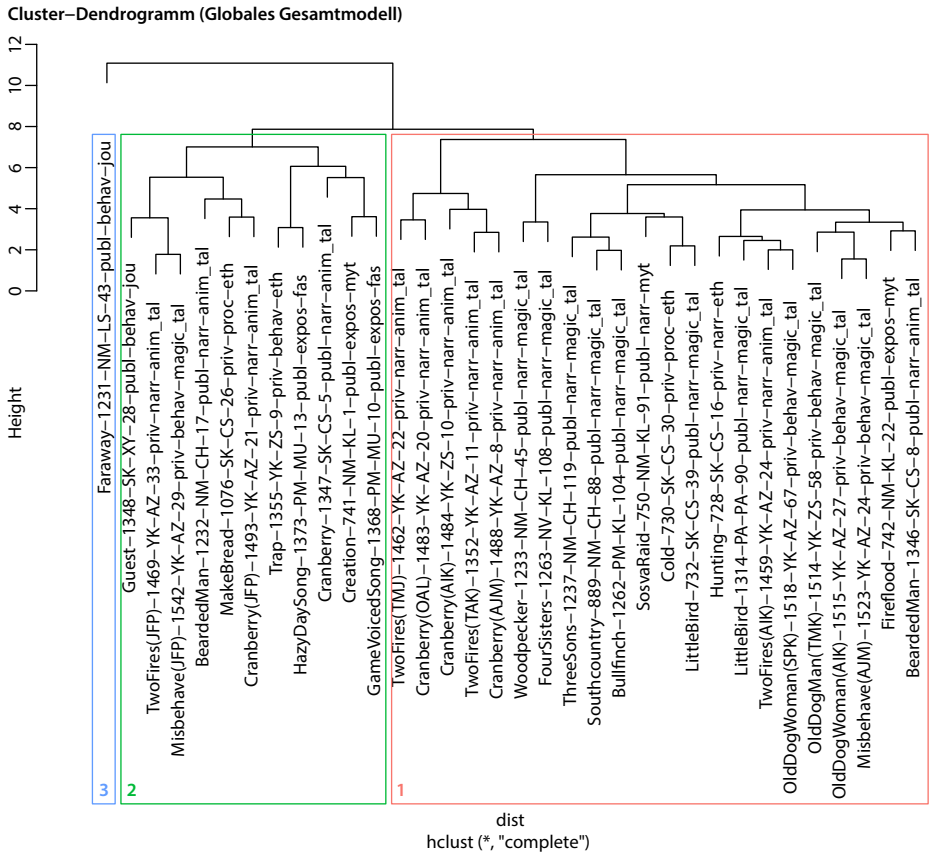


Plot C.12: Feature-Importance für COMM_SIT-Klassen (Global-relationale Parameter)

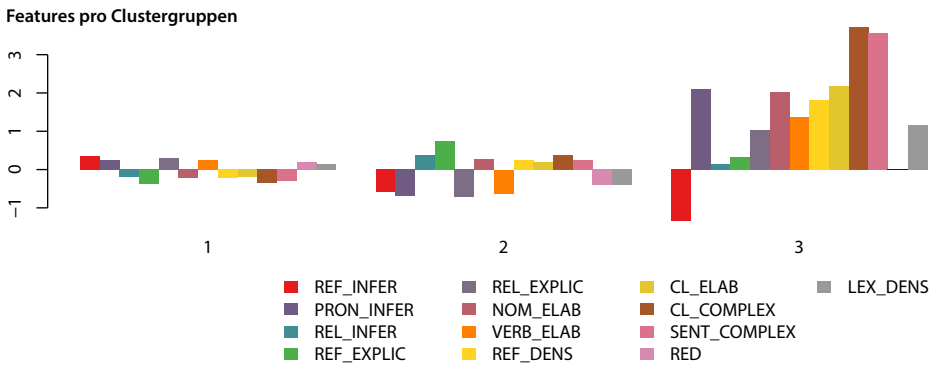


Plot C.13: Feature-Importance für DISC_STRUCT-Klassen (Global-relationale Parameter)

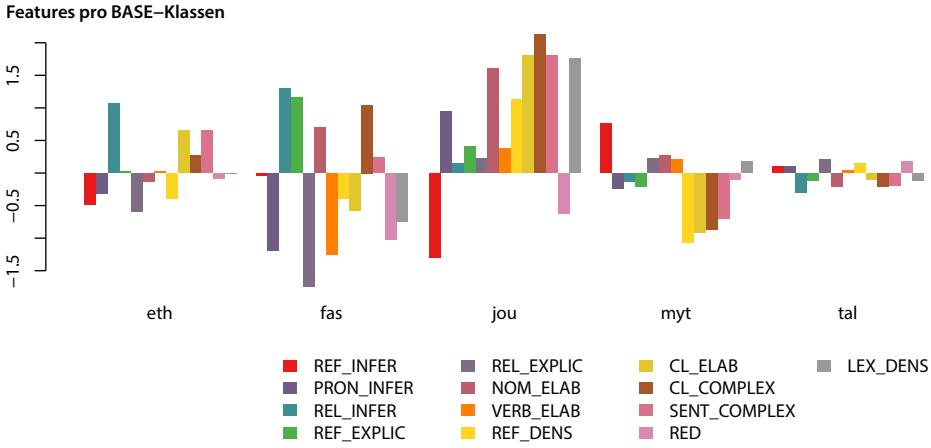
D Plots zu 6.2.4 (Globales Gesamtmodell)



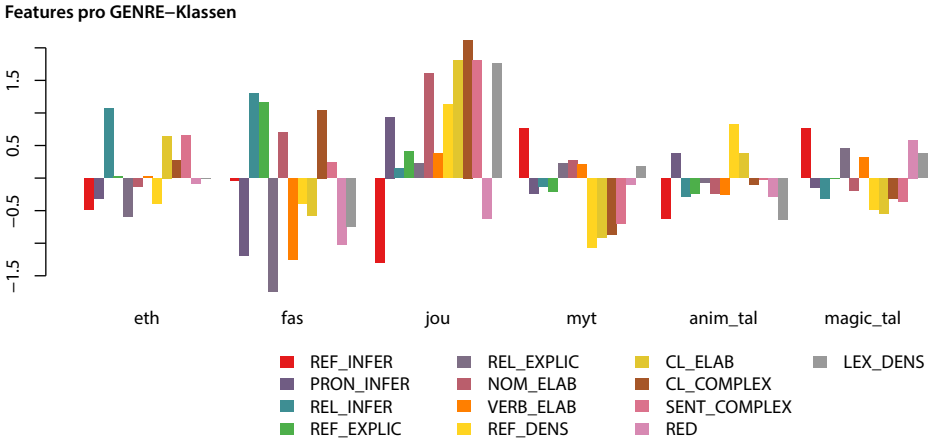
Plot D.1: Cluster-Dendrogramm (Globales Gesamtmodell)



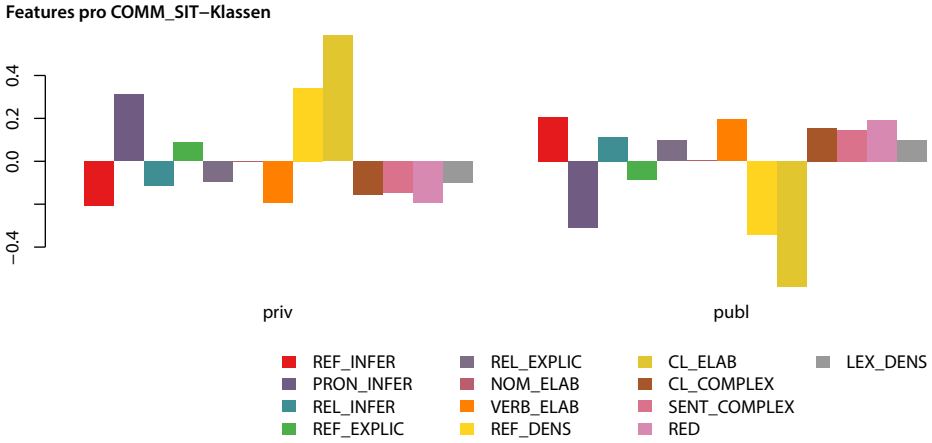
Plot D.2: Clustergruppierte Average-Scores-Barplots (Globales Gesamtmodell)



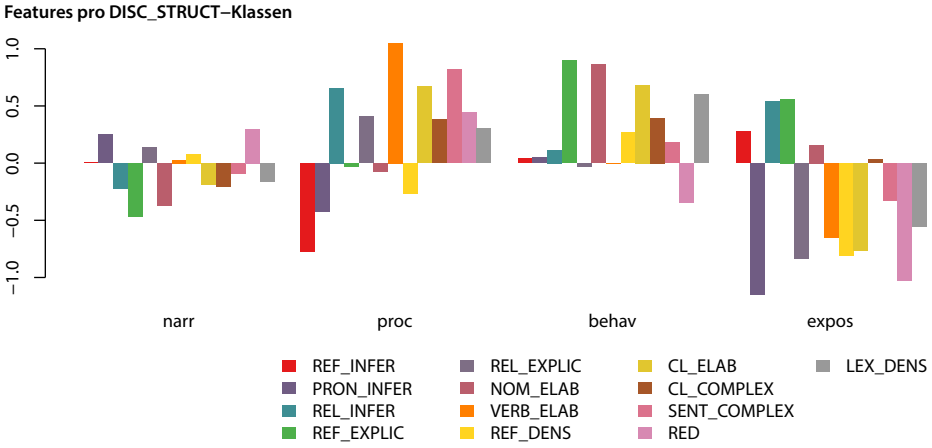
Plot D.3: BASE-gruppierte Average-Scores-Barplots (Globales Gesamtmodell)



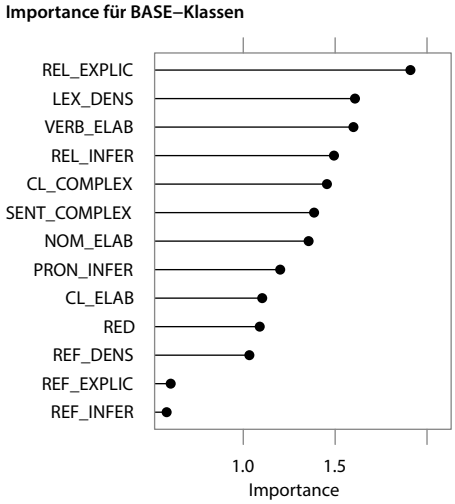
Plot D.4: GENRE-gruppierte Average-Scores-Barplots (Globales Gesamtmodell)



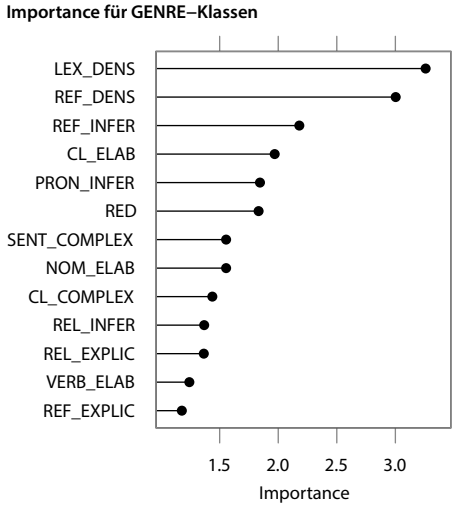
Plot D.5: COMM_SIT-gruppierte Average-Scores-Barplots (Globales Gesamtmodell)



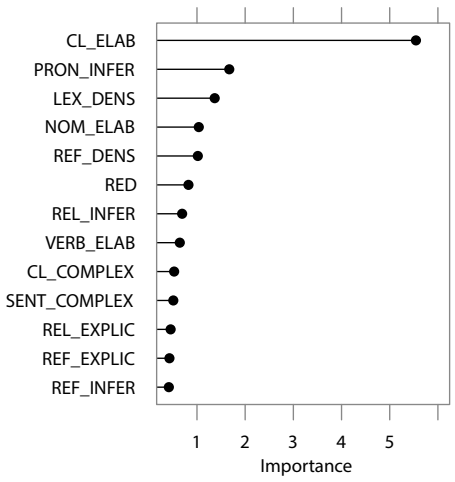
Plot D.6: DISC_STRUCT-gruppierte Average-Scores-Barplots (Globales Gesamtmodell)



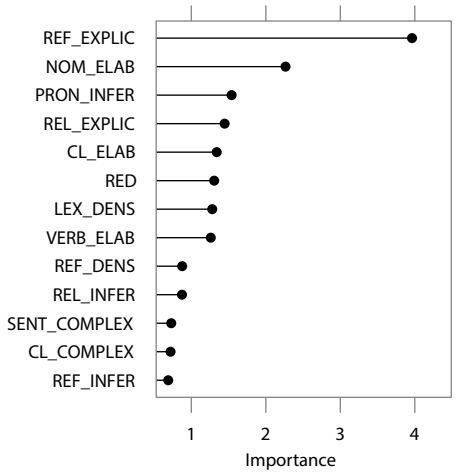
Plot D.7: Feature-Importance für BASE-Klassen (Globales Gesamtmodell)



Plot D.8: Feature-Importance für GENRE-Klassen (Globales Gesamtmodell)

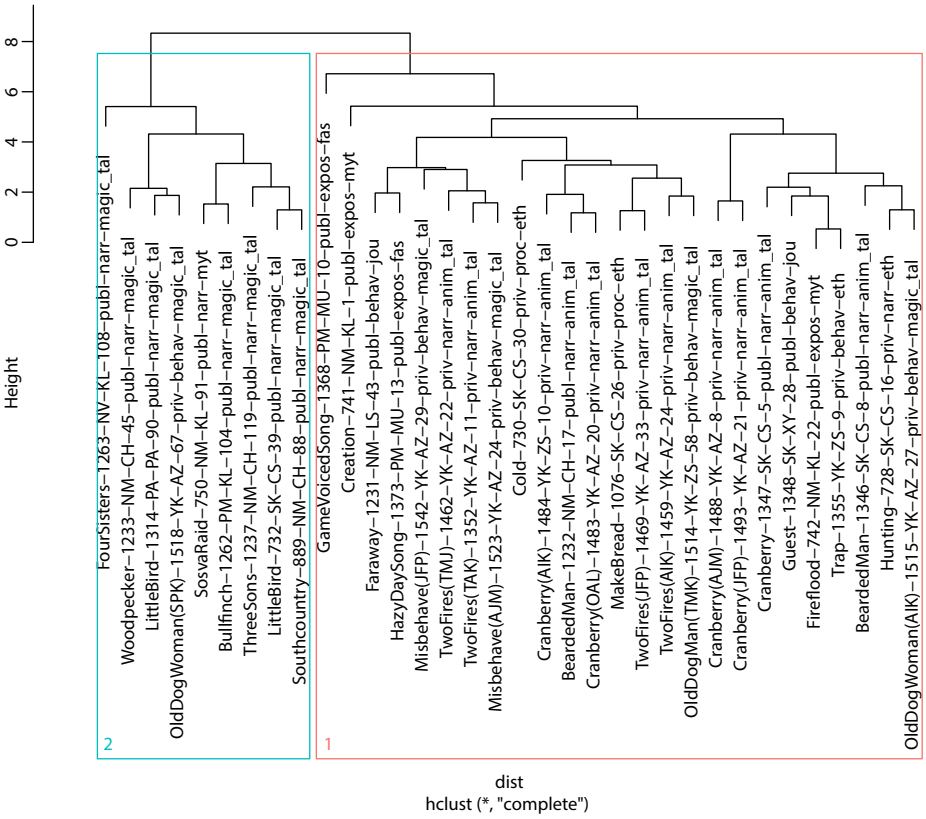


Plot D.9: Feature-Importance für COMM_SIT-Klassen (Globales Gesamtmodell)

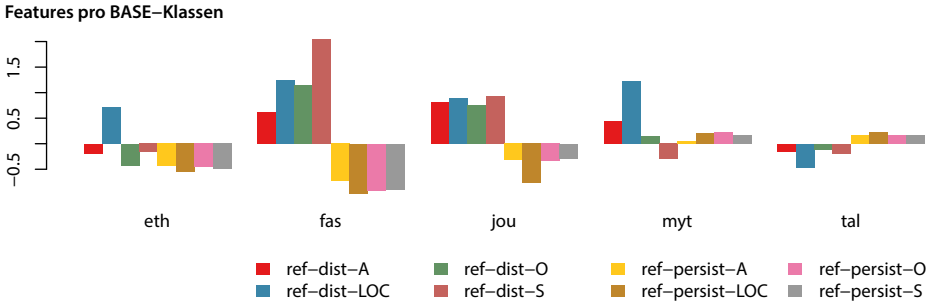


Plot D.10: Feature-Importance für DISC_STRUCT-Klassen (Globales Gesamtmodell)

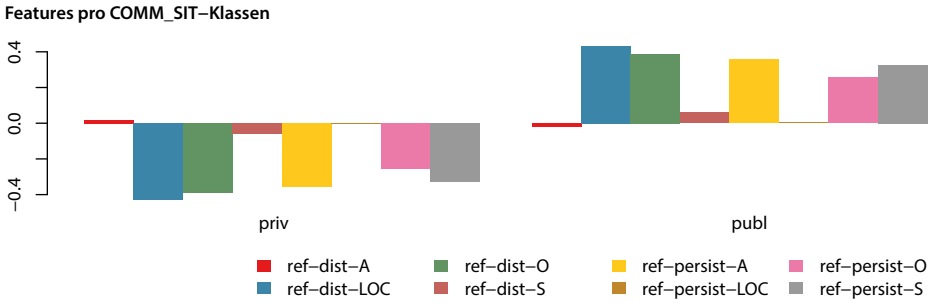
E Plots zu 6.3.3 (Distanz-Persistenz-Modell)



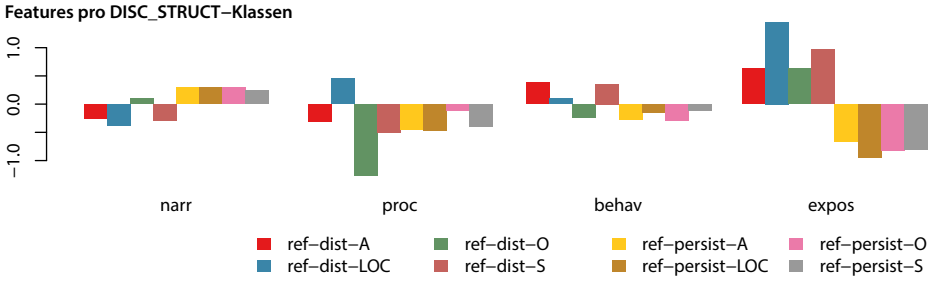
Plot E.1: Cluster-Dendrogramm (Distanz-Persistenz-Modell)



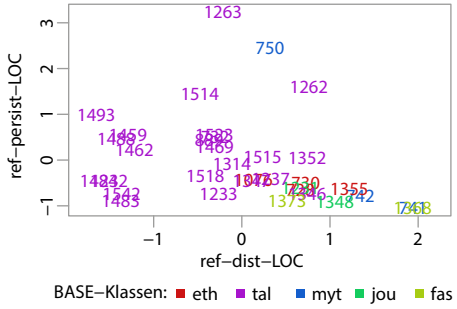
Plot E.2: BASE-gruppierte Average-Scores-Barplots (Distanz-Persistenz-Modell)



Plot E.3: COMM_SIT-gruppierte Average-Scores-Barplots (Distanz-Persistenz-Modell)

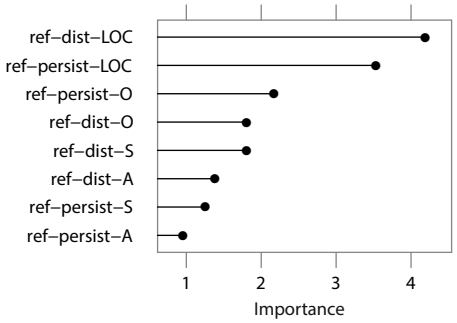


Plot E.4: DISC_STRUCT-gruppierte Average-Scores-Barplots (Distanz-Persistenz-Modell)



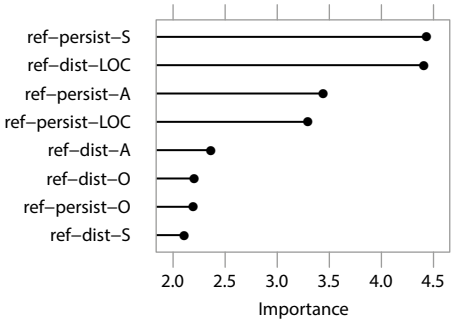
Plot E.5: Scatterplot LOC-Bereich nach BASE-Klassen (Distanz-Persistenz-Modell)

Importance für BASE−Klassen



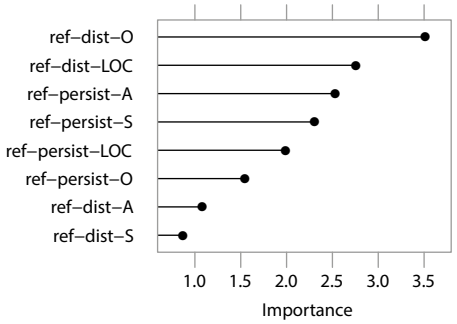
Plot E.6: Feature-Importance für BASE-Klassen (Distanz-Persistenz-Modell)

Importance für GENRE−Klassen



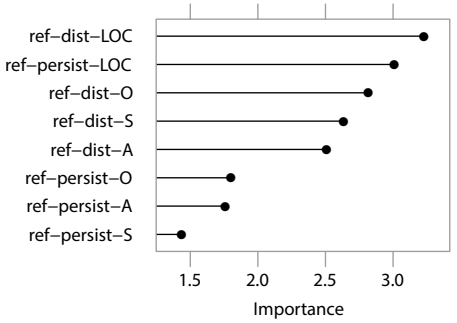
Plot E.7: Feature-Importance für GENRE-Klassen (Distanz-Persistenz-Modell)

Importance für COMM_SIT−Klassen



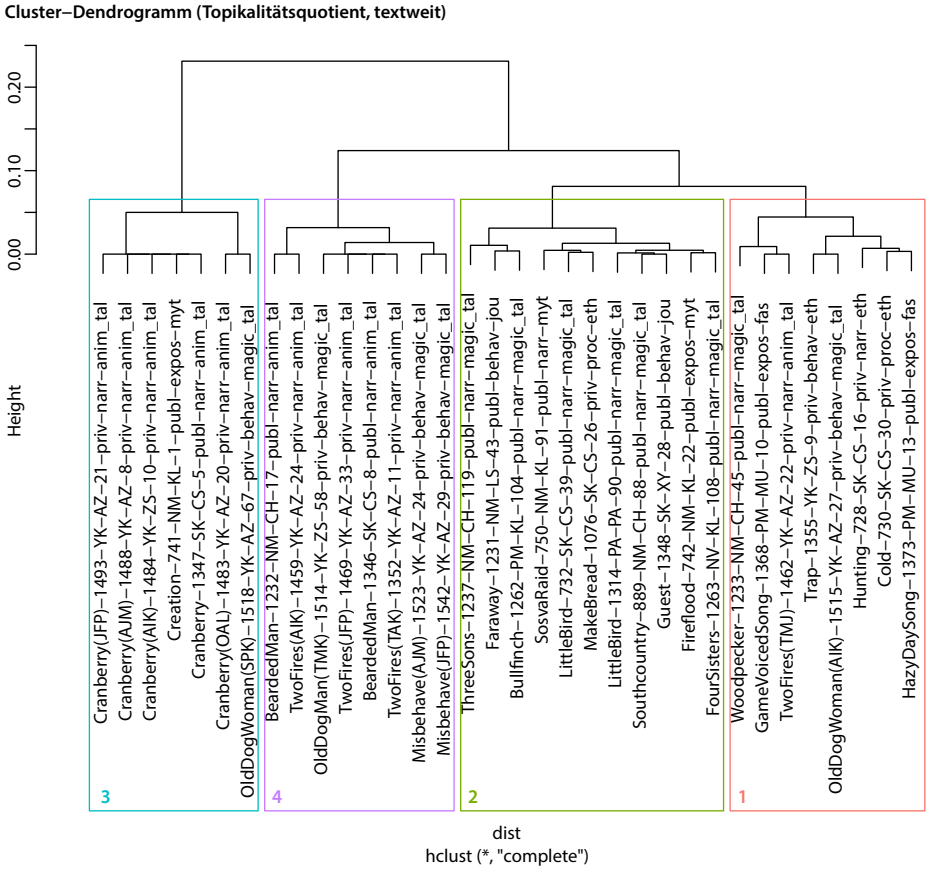
Plot E.8: Feature-Importance für COMM_SIT-Klassen (Distanz-Persistenz-Modell)

Importance für DISC_STRUCT−Klassen

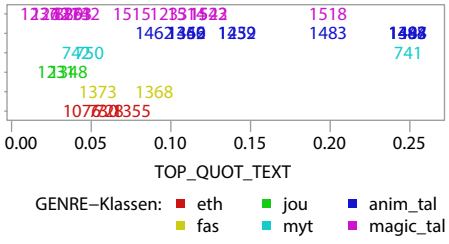


Plot E.9: Feature-Importance für DISC_STRUCT-Klassen (Distanz-Persistenz-Modell)

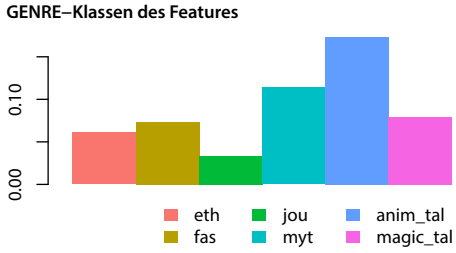
F Plots zu 6.3.4 (Textweiter Topikalitätsquotient)



Plot F.1: Cluster-Dendrogramm (Textweiter Topikalitätsquotient)

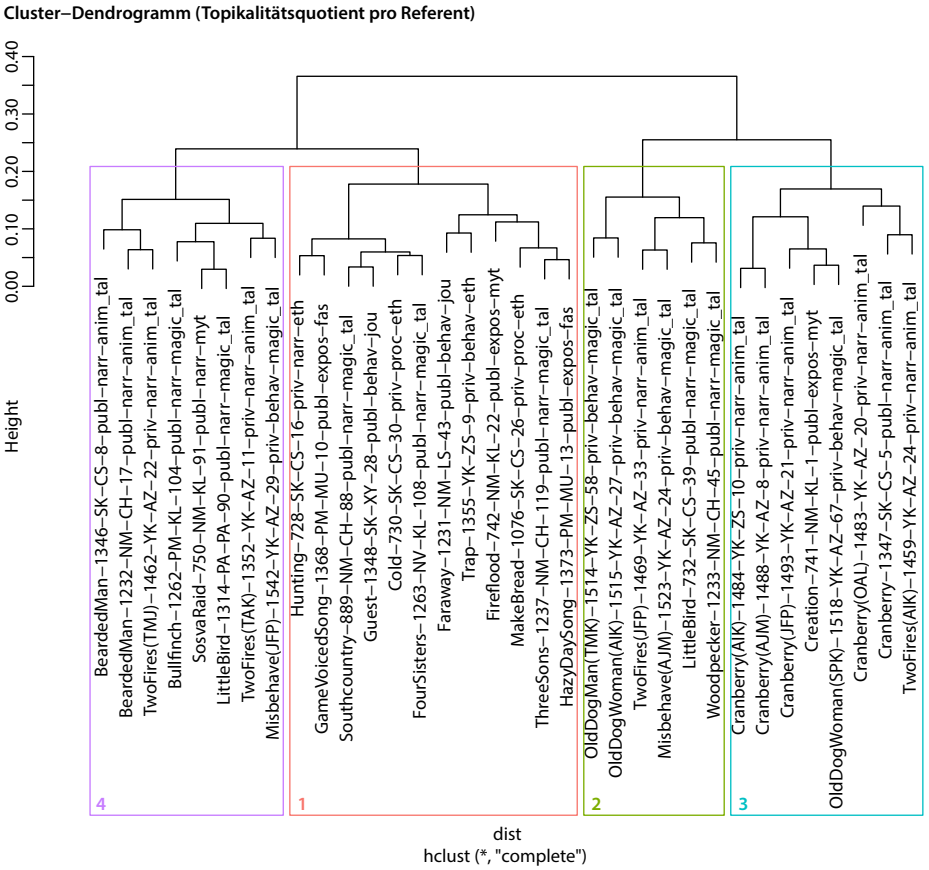


Plot F.2: Scatterplot nach GENRE-Klassen (Textweiter Topikalitätsquotient)

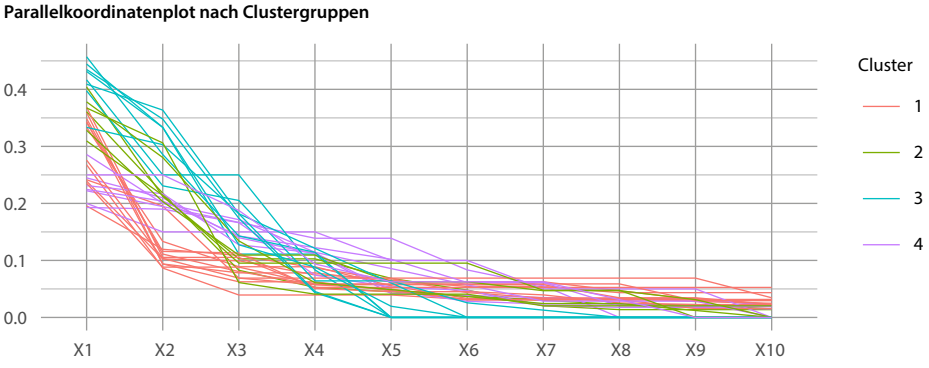


Plot F.3: GENRE-gruppierte Average-Scores-Barplots (Textweiter Topikalitätsquotient)

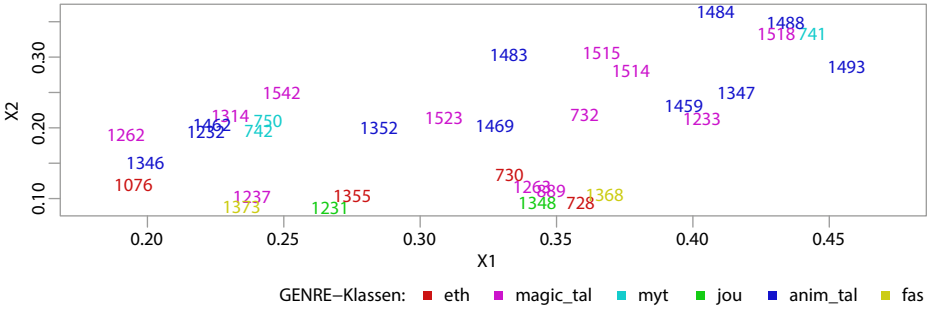
G Plots zu 6.3.5 (Topikalitätsquotienten-Verteilung)



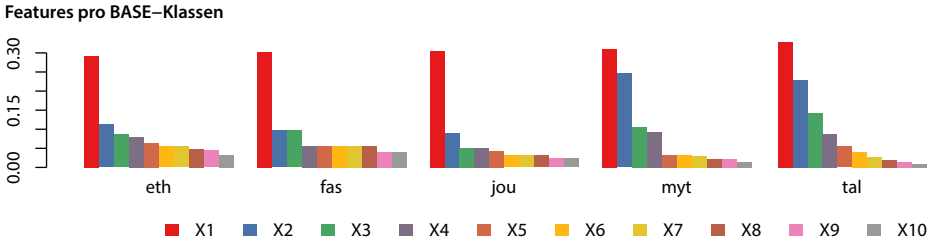
Plot G.1: Cluster-Dendrogramm (Topikalitätsquotienten)



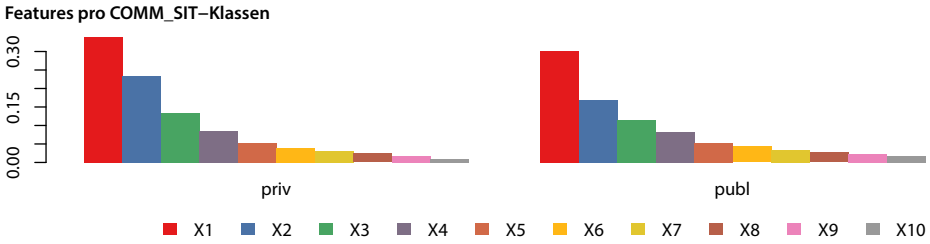
Plot G.2: Parallelkoordinatenplot nach Clustergruppen (Topikalitätsquotienten)



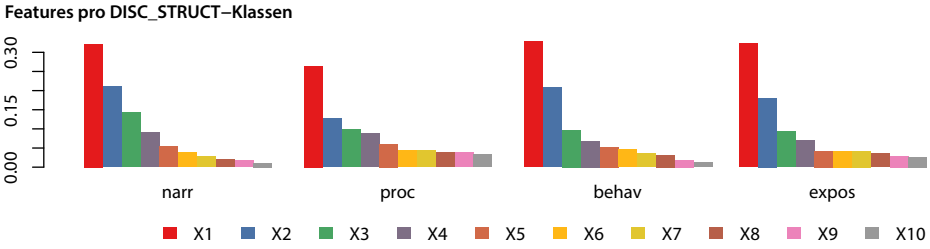
Plot G.3: Topikalitätsstärke erster und zweiter Referent nach GENRE-Klassen (Topikalitätsquotienten)



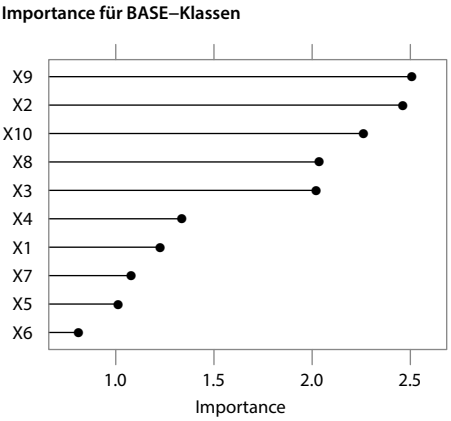
Plot G.4: BASE-gruppierte Average-Scores-Barplots (Topikalitätsquotienten)



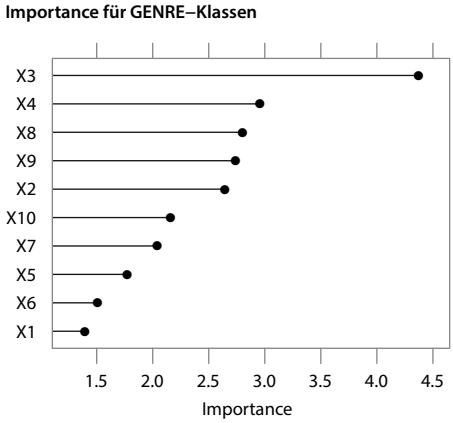
Plot G.5: COMM_SIT-gruppierte Average-Scores-Barplots (Topikalitätsquotienten)



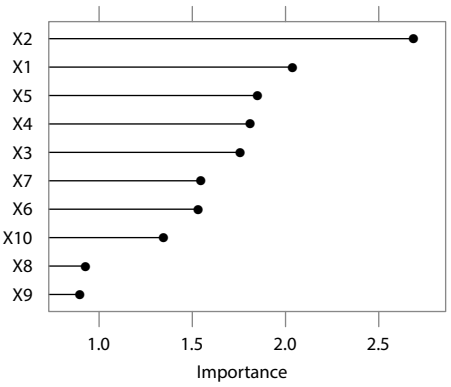
Plot G.6: DISC_STRUCT-gruppierte Average-Scores-Barplots (Topikalitätsquotienten)



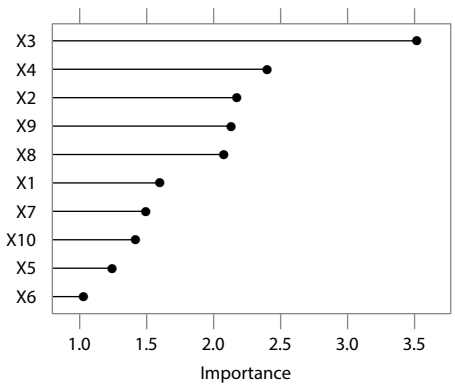
Plot G.7: Feature-Importance für BASE-Klassen (Topikalitätsquotienten)



Plot G.8: Feature-Importance für GENRE-Klassen (Topikalitätsquotienten)

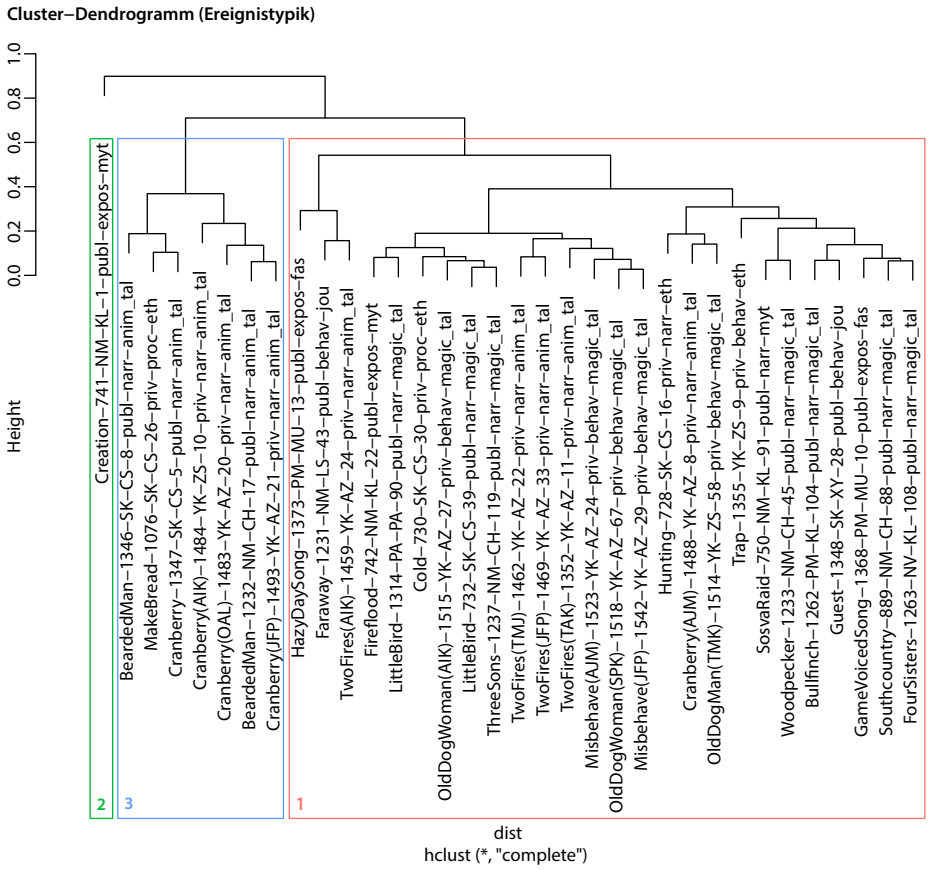


Plot G.9: Feature-Importance für COMM_SIT-Klassen (Topikalitätsquotienten)

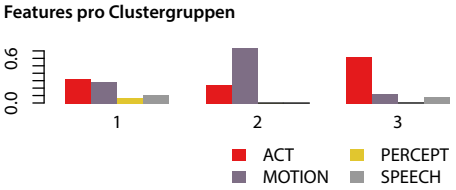


Plot G.10: Feature-Importance für DISC_STRUCT-Klassen (Topikalitätsquotienten)

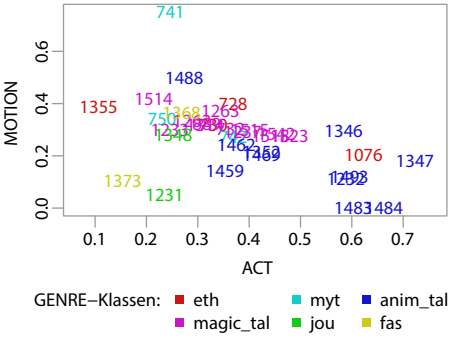
H Plots zu 6.4.1 (Ereignistypik)



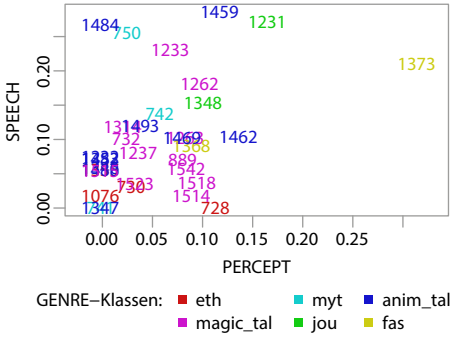
Plot H.1: Cluster-Dendrogramm (Ereignistypik)



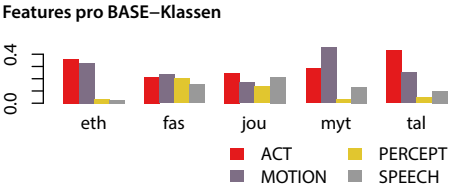
Plot H.2: Clustergruppierte Average-Scores-Barplots (Ereignistypik)



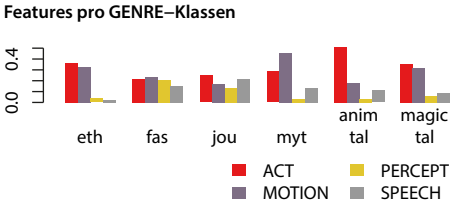
Plot H.3: Handlungs- und bewegungsbezogene Ereignistypen nach GENRE-Klassifizierung (Ereignistypik)



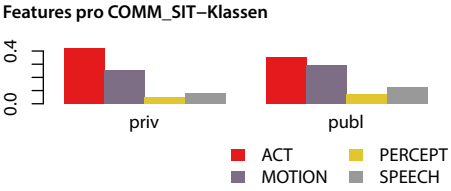
Plot H.4: Wahrnehmungs- und sprachbezogene Ereignistypen nach GENRE-Klassifizierung (Ereignistypik)



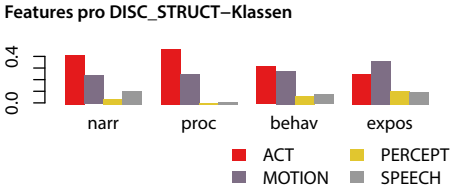
Plot H.5: BASE-gruppierte Average-Scores-Barplots (Ereignistypik)



Plot H.6: GENRE-gruppierte Average-Scores-Barplots (Ereignistypik)

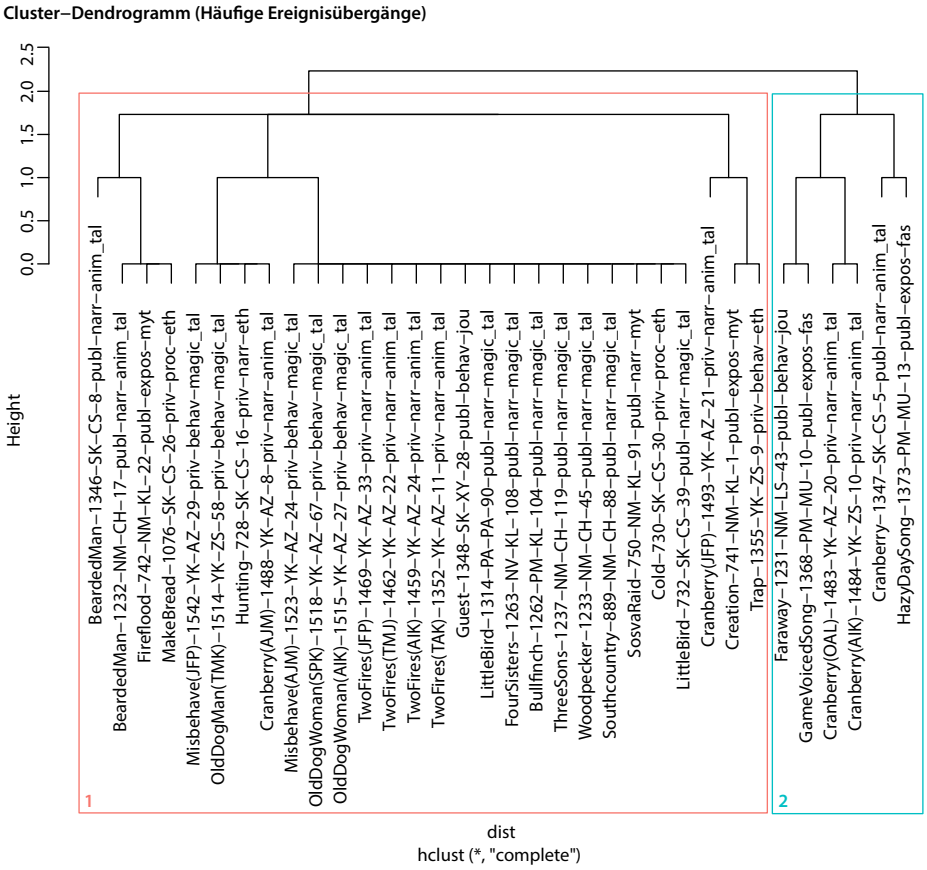


Plot H.7: COMM_SIT-gruppierte Average-Scores-Barplots (Ereignistypik)

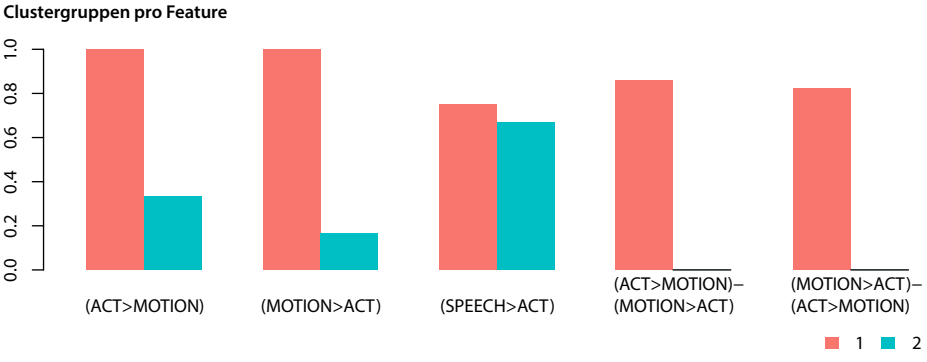


Plot H.8: DISC_STRUCT-gruppierte Average-Scores-Barplots (Ereignistypik)

I Plots zu 6.4.2 (Häufige Ereignisübergänge)

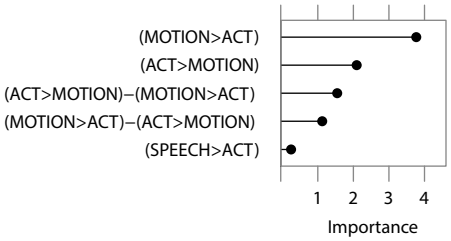


Plot I.1: Cluster-Dendrogramm (Häufige Ereignisübergänge)



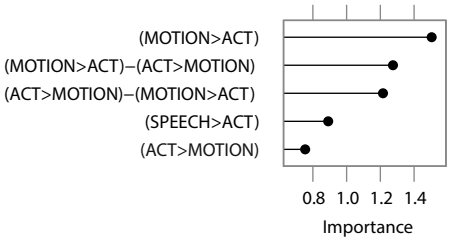
Plot I.2: Featuregruppierte Average-Scores-Barplots der Clustergruppen (Häufige Ereignisübergänge)

Importance für Clustergruppen



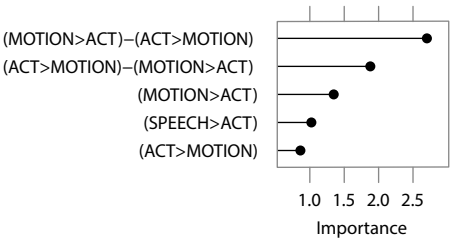
Plot I.3: Feature-Importance für Clustergruppen (Häufige Ereignisübergänge)

Importance für BASE–Klassen



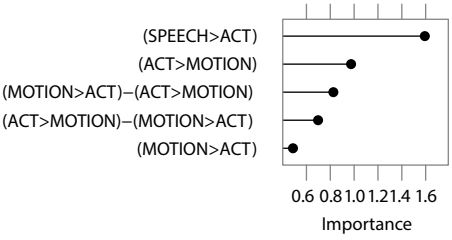
Plot I.4: Feature-Importance für BASE-Klassen (Häufige Ereignisübergänge)

Importance für GENRE–Klassen



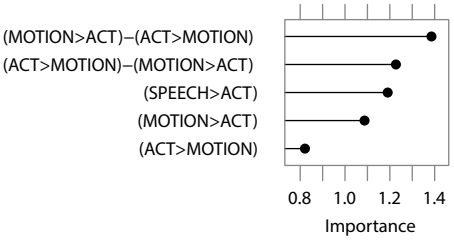
Plot I.5: Feature-Importance für GENRE-Klassen (Häufige Ereignisübergänge)

Importance für COMM_SIT–Klassen



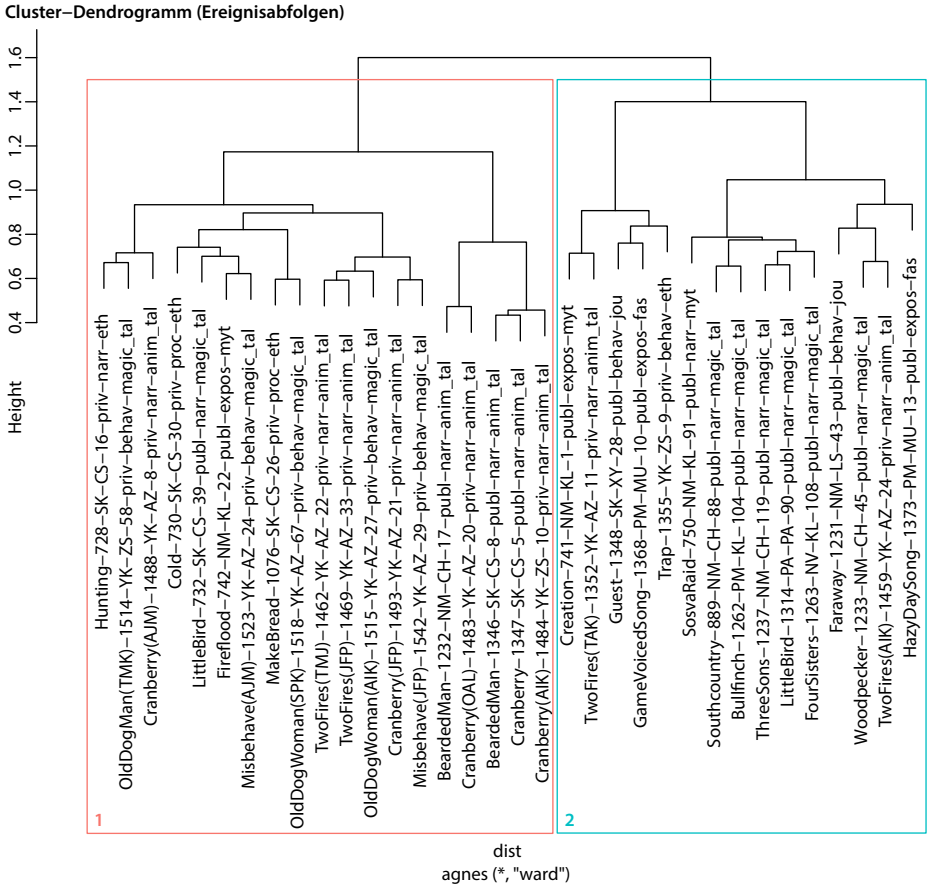
Plot I.6: Feature-Importance für COMM_SIT-Klassen (Häufige Ereignisübergänge)

Importance für DISC_STRUCT–Klassen

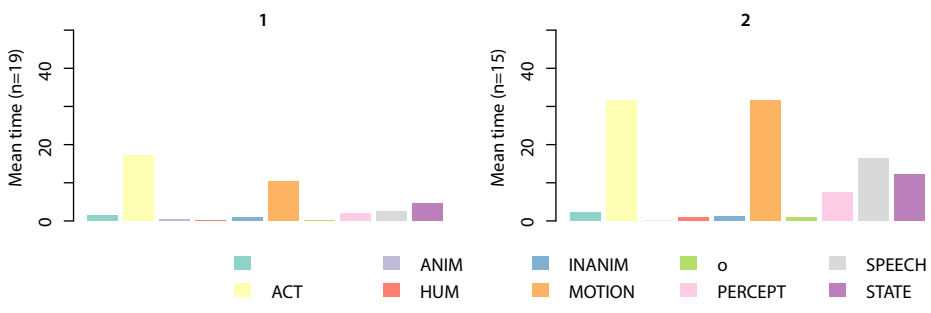


Plot I.7: Feature-Importance für DISC_STRUCT-Klassen (Häufige Ereignisübergänge)

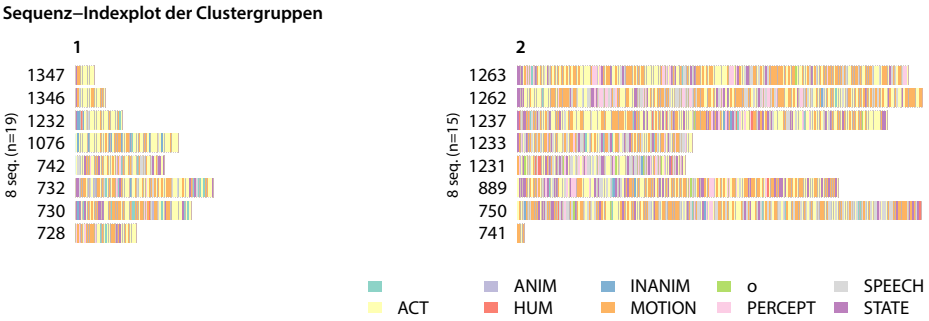
J Plots zu 6.4.3 (Globale Ereignisabfolge)



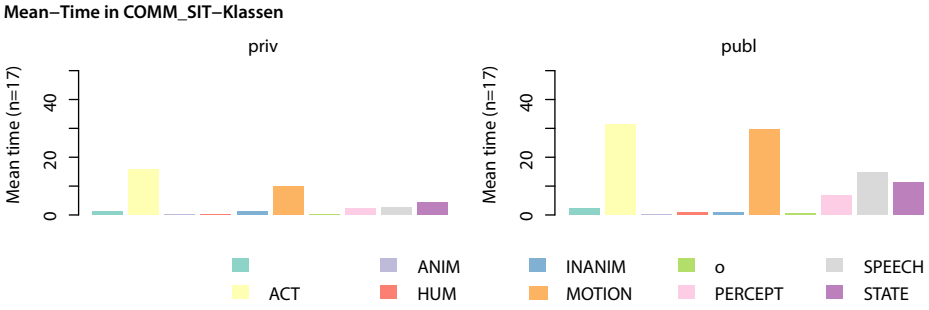
Plot J.1: Cluster-Dendrogramm (Globale Abfolge Ereigniszustände)



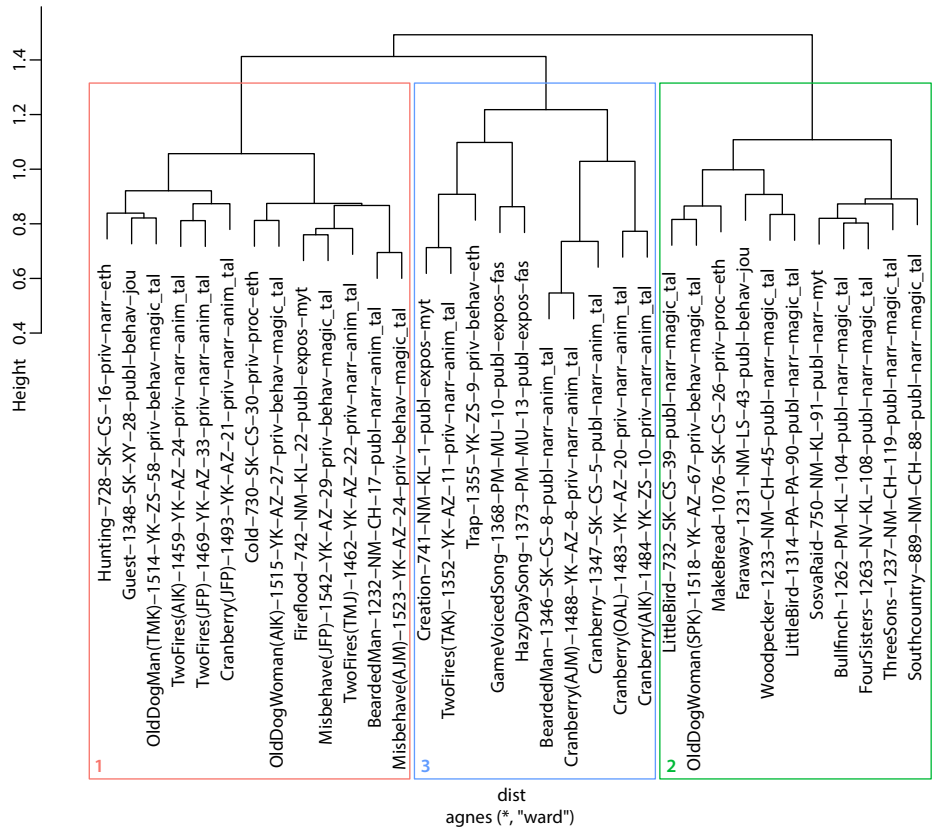
Plot J.2: Durchschnittliche Verweildauer in Clustergruppen (Globale Abfolge Ereigniszustände)



Plot J.3: Sequenz-Indexplot für Clustergruppen (Globale Abfolge Ereigniszustände)

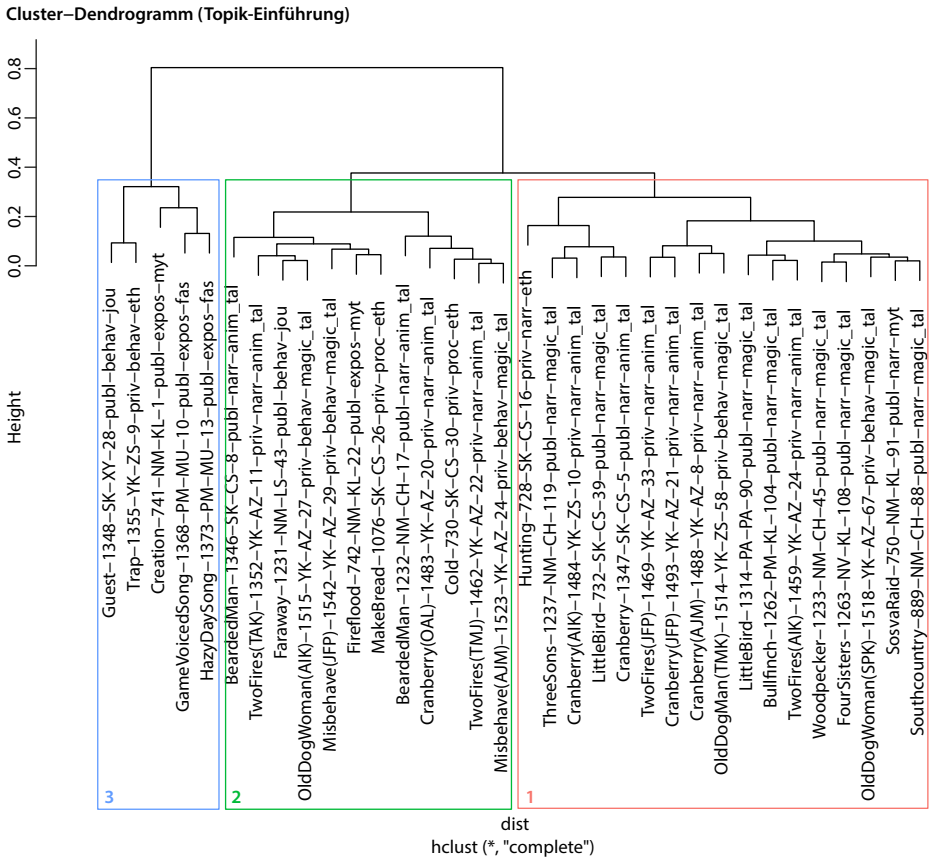


Plot J.4: Durchschnittliche Verweildauer in COMM_SIT-Klassen (Globale Abfolge Ereigniszustände)

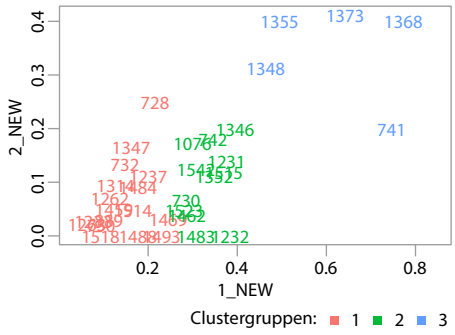


Plot J.5: Cluster-Dendrogramm (Globale Abfolge Ereignisübergänge)

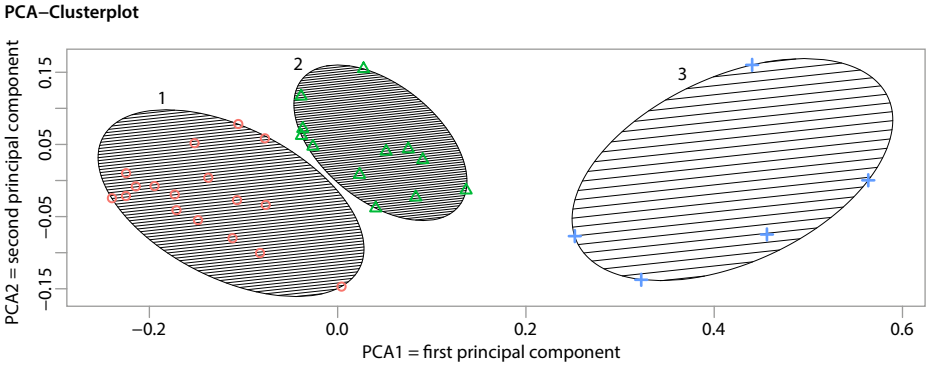
K Plots zu 6.5.1 (Topik-Einführungen)



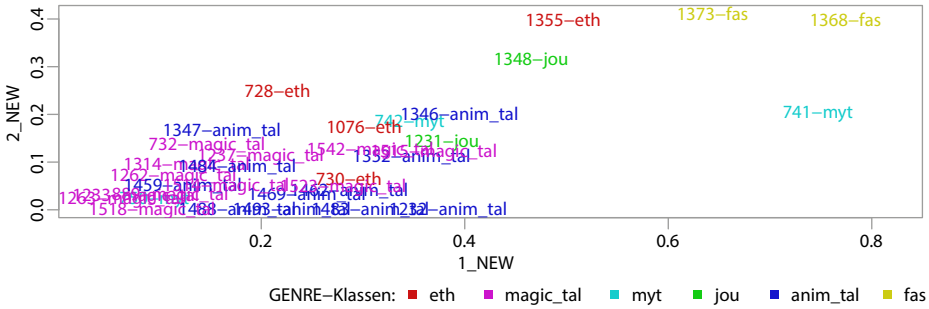
Plot K.1: Cluster-Dendrogramm (Topik-Einführungen)



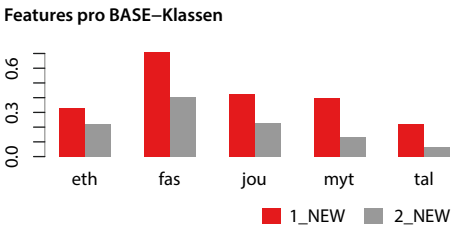
Plot K.2: Scatterplot nach Clustergruppen (Topik-Einführungen)



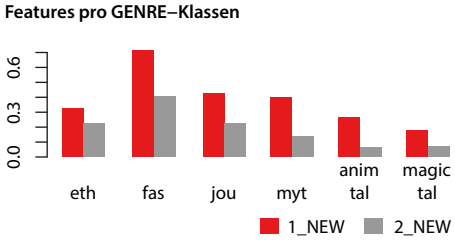
Plot K.3: Hauptkomponenten-Clusterplot (Topik-Einführungen)



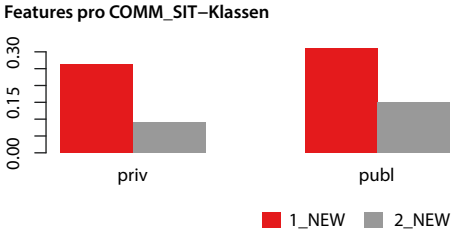
Plot K.4: Scatterplot nach GENRE-Klassen (Topik-Einführungen)



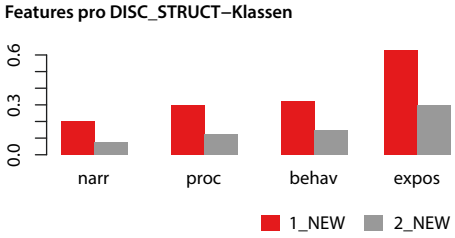
Plot K.5: BASE-gruppierte Average-Scores-Barplots (Topik-Einführungen)



Plot K.6: GENRE-gruppierte Average-Scores-Barplots (Topik-Einführungen)

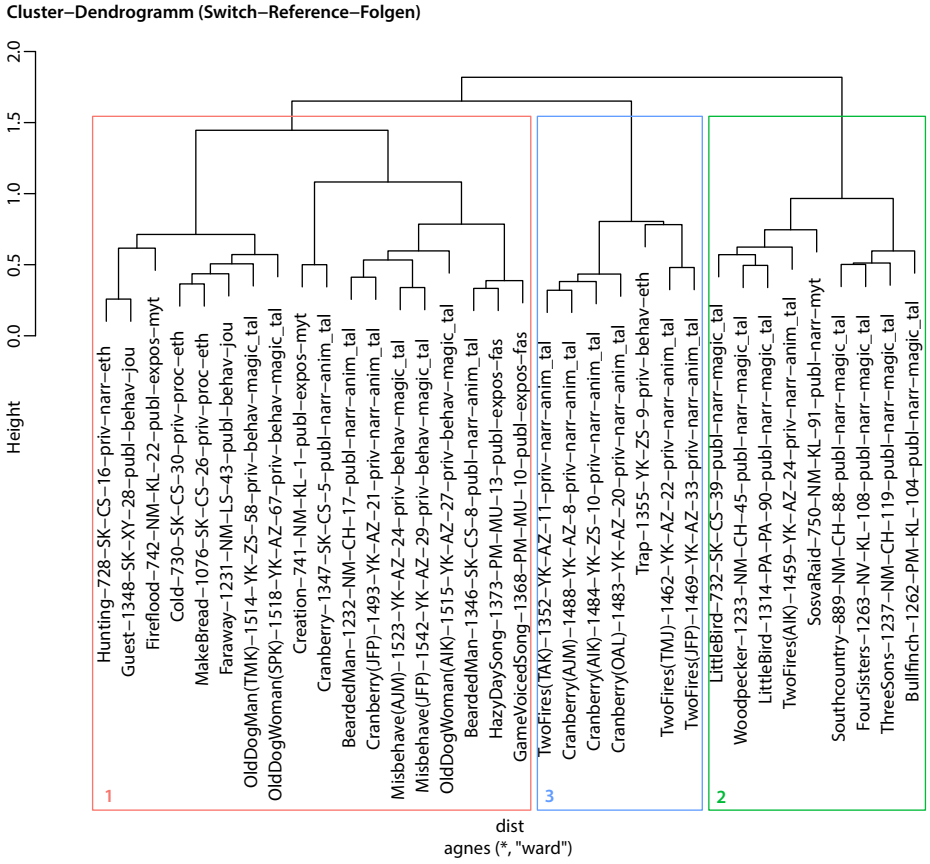


Plot K.7: COMM_SIT-gruppierte Average-Scores-Barplots (Topik-Einführungen)

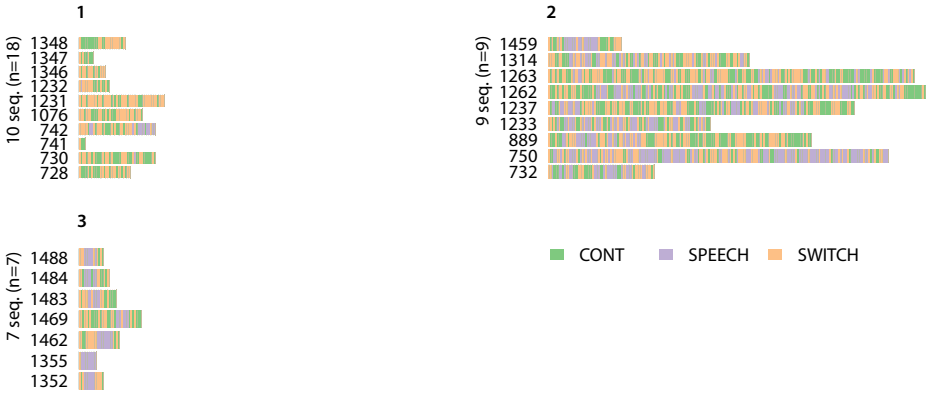


Plot K.8: DISC_STRUCT-gruppierte Average-Scores-Barplots (Topik-Einführungen)

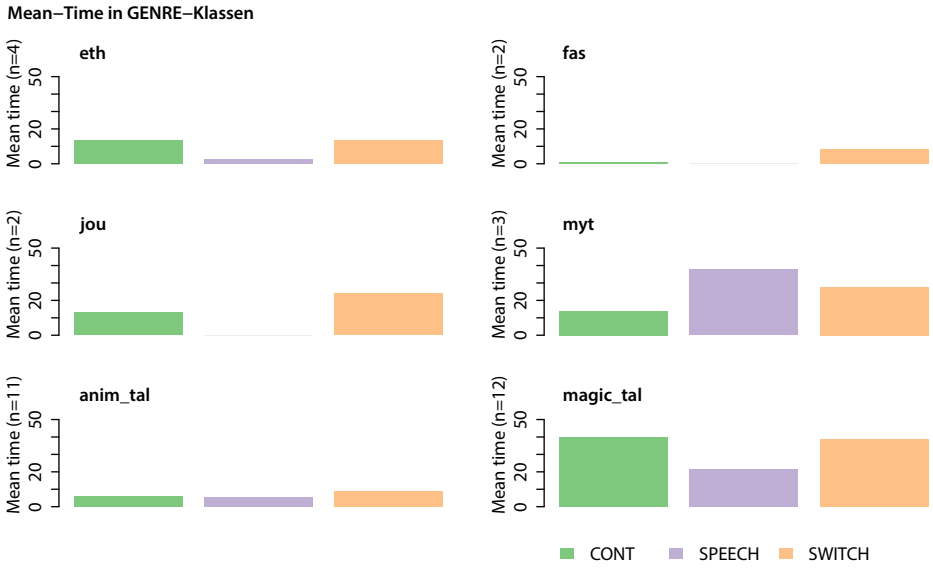
L Plots zu 6.5.2 (Switch-Reference-Sequenzen)



Plot L.1: Cluster-Dendrogramm (Switch-Reference-Sequenzen)



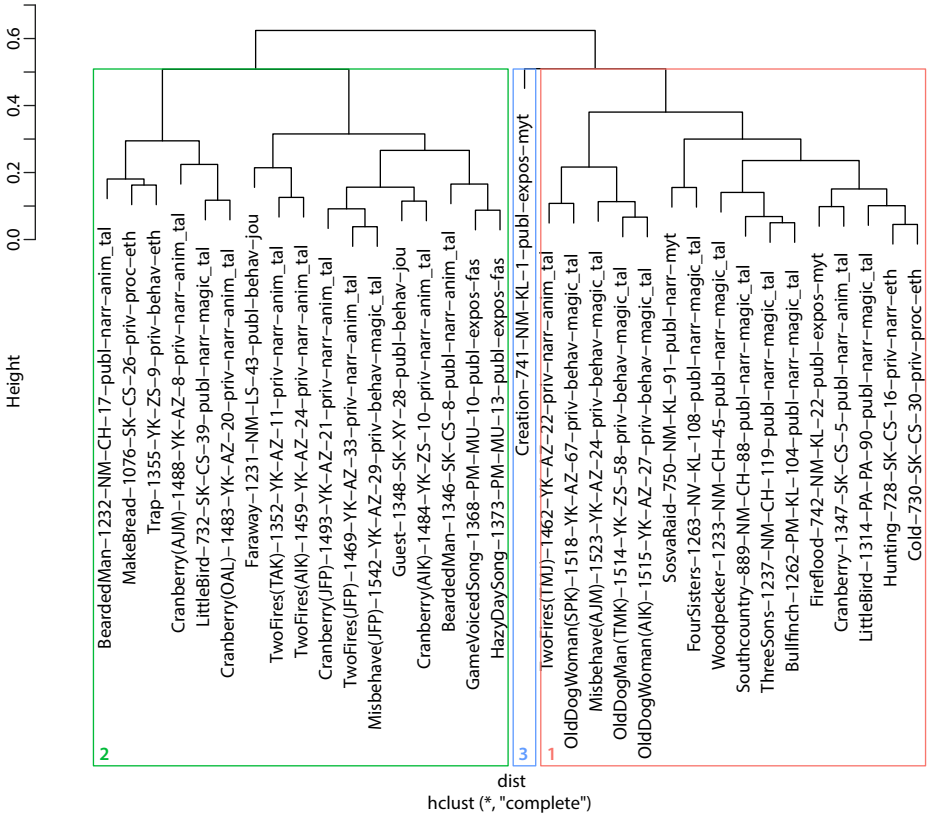
Plot L.2: Sequenz-Indexplot für Clustergruppen (Switch-Reference-Sequenzen)



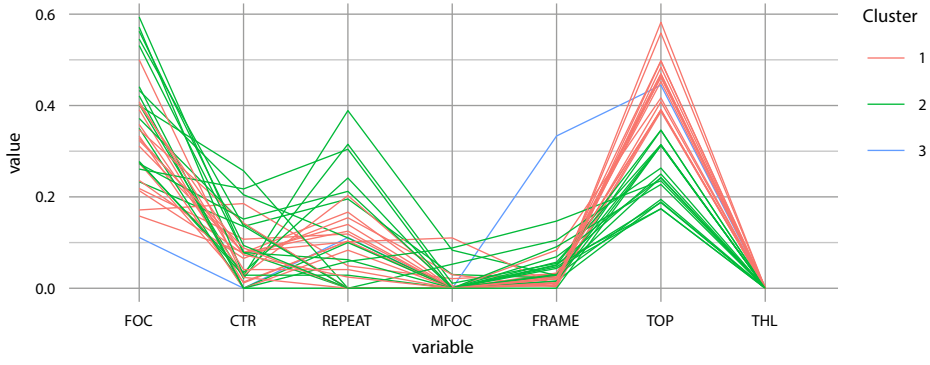
Plot L.3: Durchschnittliche Verweildauer in GENRE-Klassen (Switch-Reference-Sequenzen)

M Plots zu 6.5.3 (Fokussierungstypik)

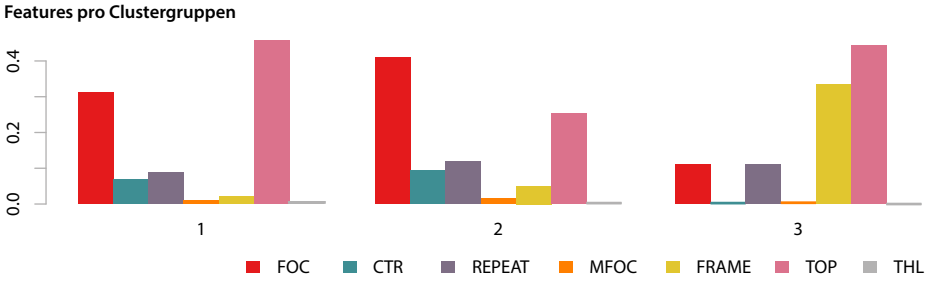
Cluster−Dendrogramm (Fokussierungstypik)



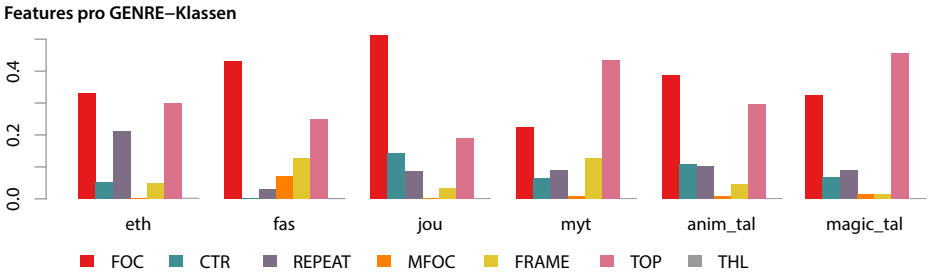
Plot M.1: Cluster-Dendrogramm (Fokussierungstypik)



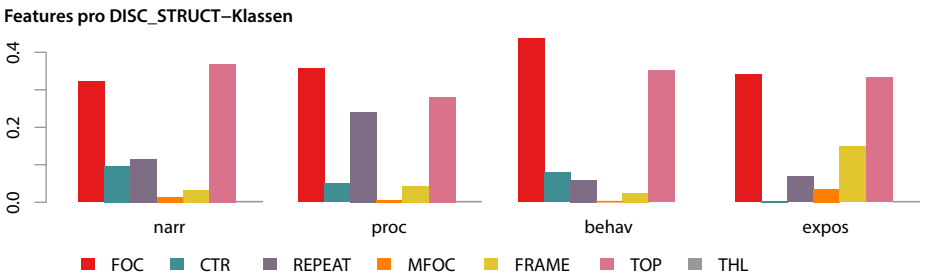
Plot M.2: Parallelkoordinatenplot nach Clustergruppen (Fokussierungstypik)



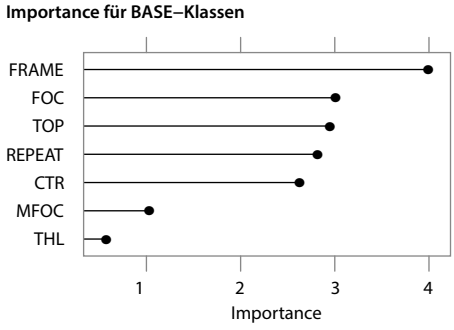
Plot M.3: Clustergruppierte Average-Scores-Barplots (Fokussierungstypik)



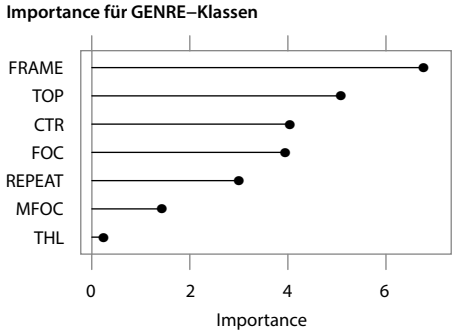
Plot M.4: GENRE-gruppierte Average-Scores-Barplots (Fokussierungstypik)



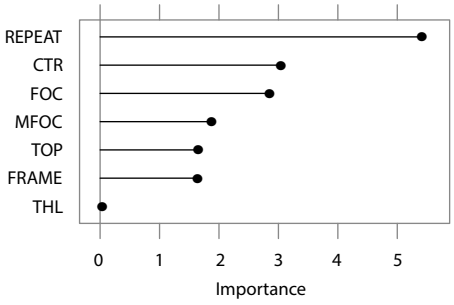
Plot M.5: DISC_STRUCT-gruppierte Average-Scores-Barplots (Fokussierungstypik)



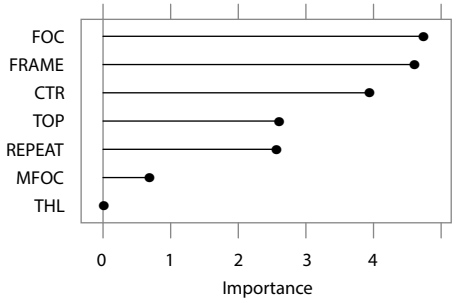
Plot M.6: Feature-Importance für BASE-Klassen (Fokussierungstypik)



Plot M.7: Feature-Importance für GENRE-Klassen (Fokussierungstypik)



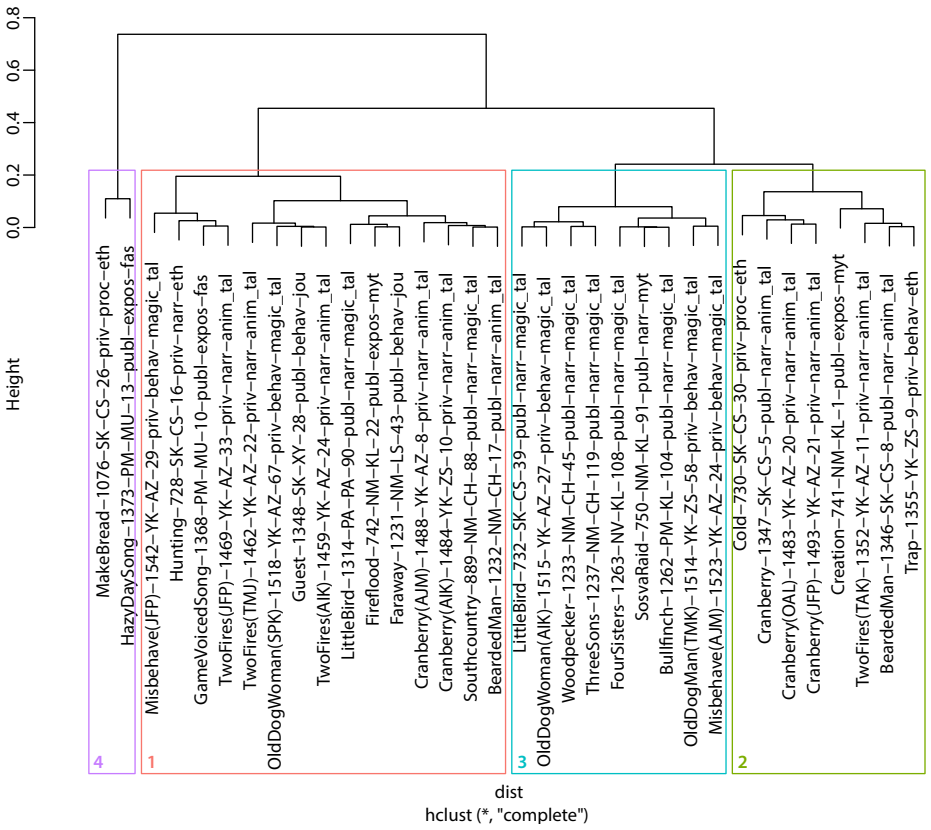
Plot M.8: Feature-Importance für COMM_SIT-Klassen (Fokussierungstypik)



Plot M.9: Feature-Importance für DISC_STRUCT-Klassen (Fokussierungstypik)

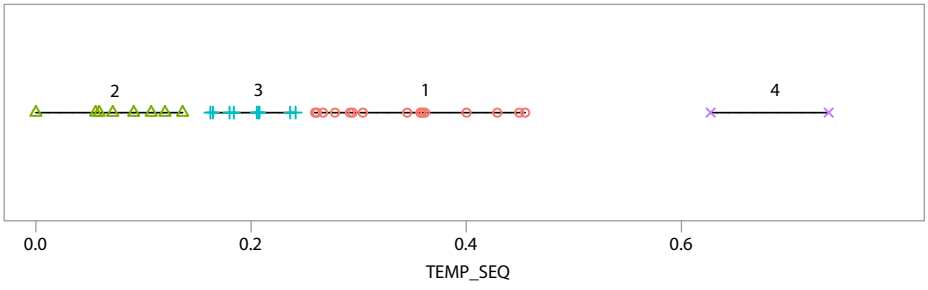
N Plots zu 6.5.4 (Temporal-Sequencing)

Cluster–Dendrogramm (Temporal-Sequencing)

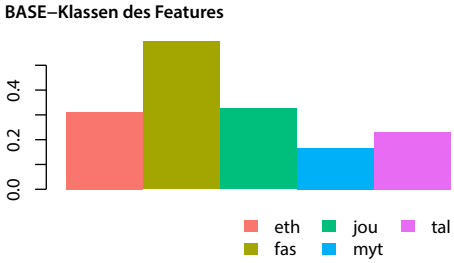


Plot N.1: Cluster-Dendrogramm (Temporal-Sequencing)

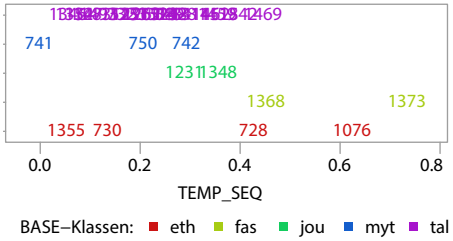
Clusterplot



Plot N.2: Clusterplot (Temporal-Sequencing)

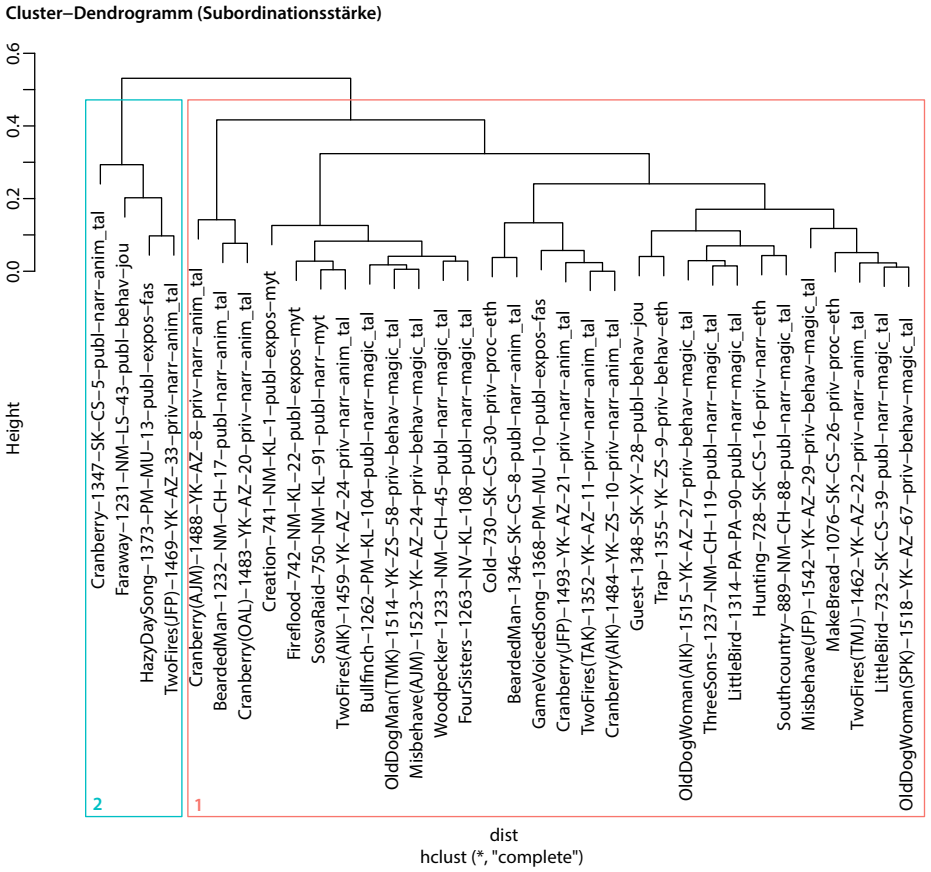


Plot N.3: BASE-gruppierte Average-Scores-Barplots (Temporal-Sequencing)

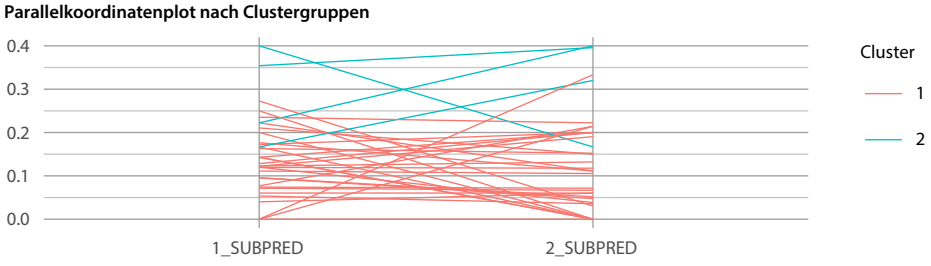


Plot N.4: Scatterplot nach BASE-Klassen (Temporal-Sequencing)

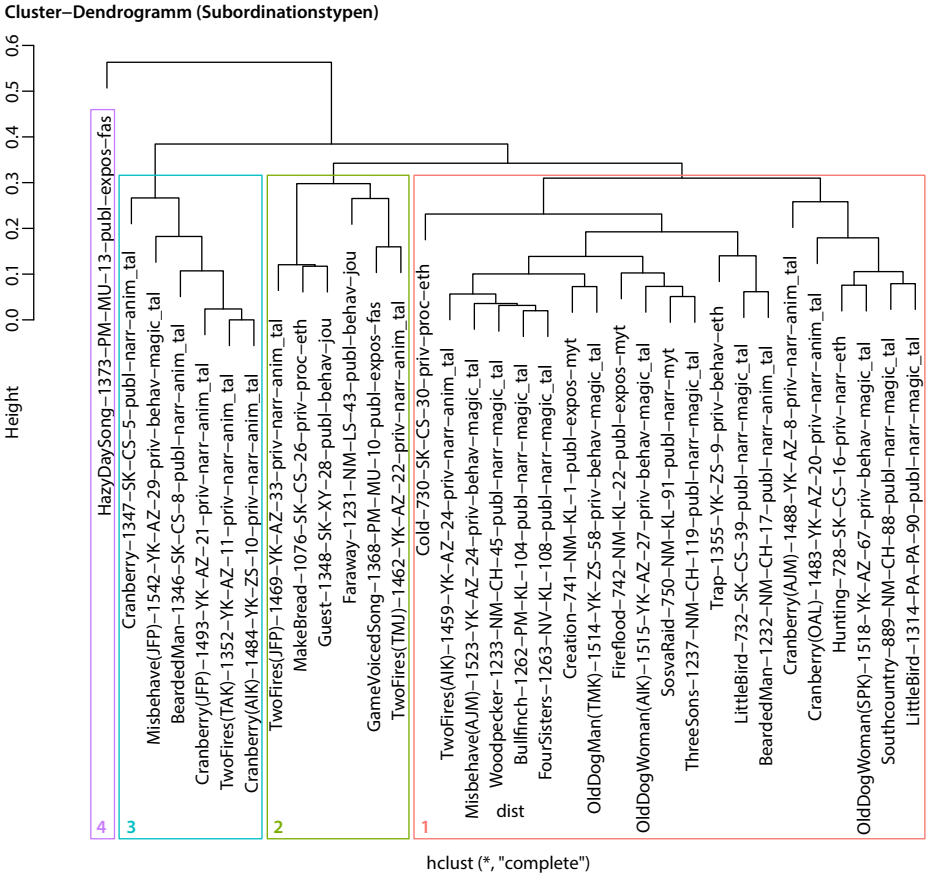
O Plots zu 6.5.5 (Komplexitätsverlauf)



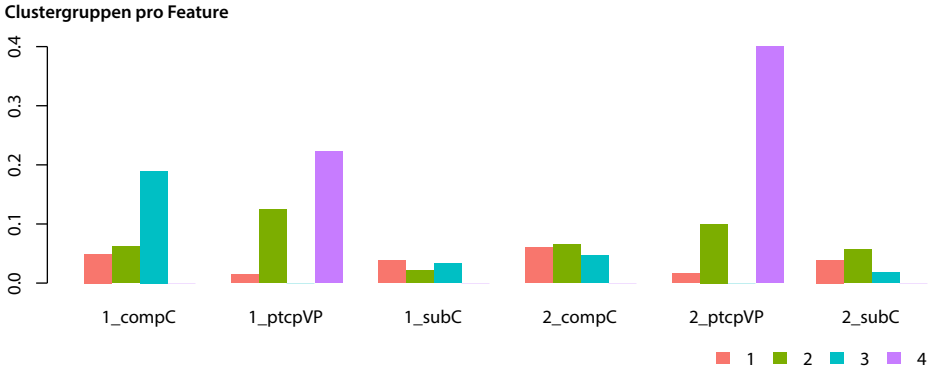
Plot O.1: Cluster-Dendrogramm (Subordinationsstärke)



Plot O.2: Parallelkoordinatenplot nach Clustergruppen (Subordinationsstärke)

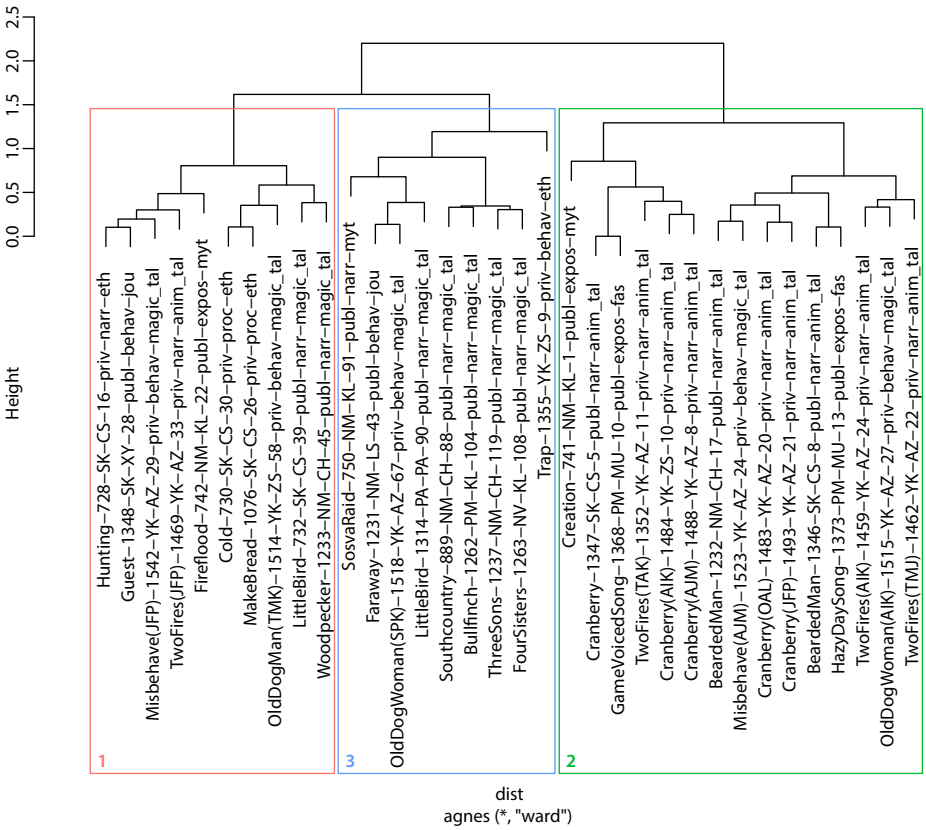


Plot O.3: Cluster-Dendrogramm (Subordinationstypen)



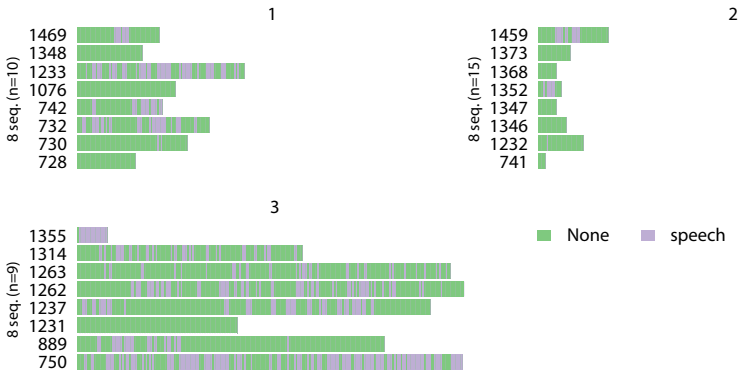
Plot O.4: Featuregruppierte Average-Scores-Barplots der Clustergruppen (Subordinationstypen)

P Plots zu 6.5.6 (Diskursstrukturelle Sequenzen)



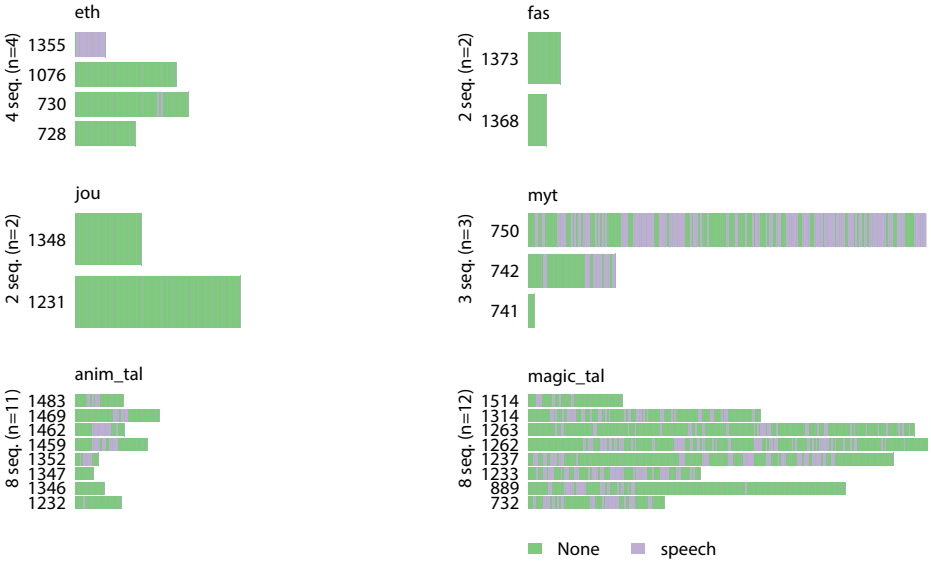
Plot P.1: Cluster-Dendrogramm (Diskursstrukturelle Sequenzen)

Sequenz–Indexplot der Clustergruppen

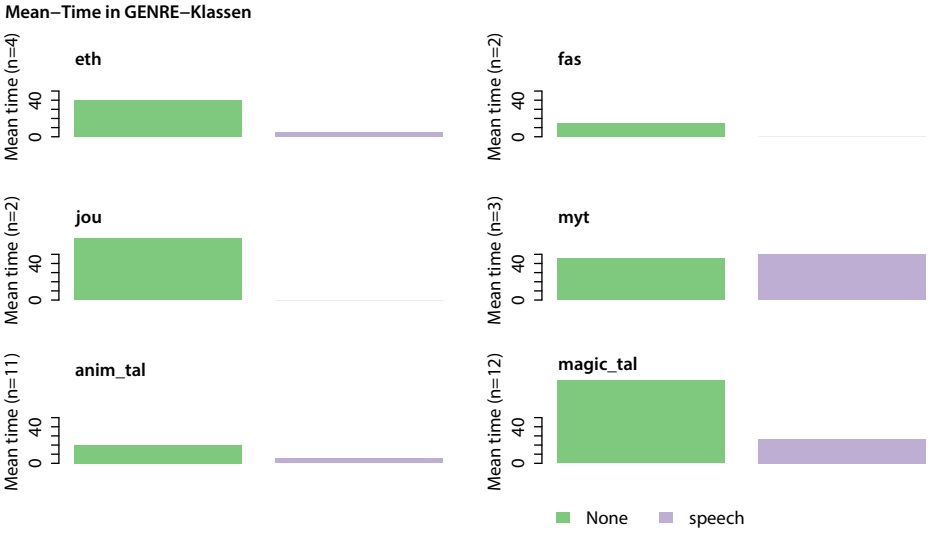


Plot P.2: Sequenz-Indexplot für Clustergruppen (Diskursstrukturelle Sequenzen)

Sequenz–Indexplot der GENRE–Klassen

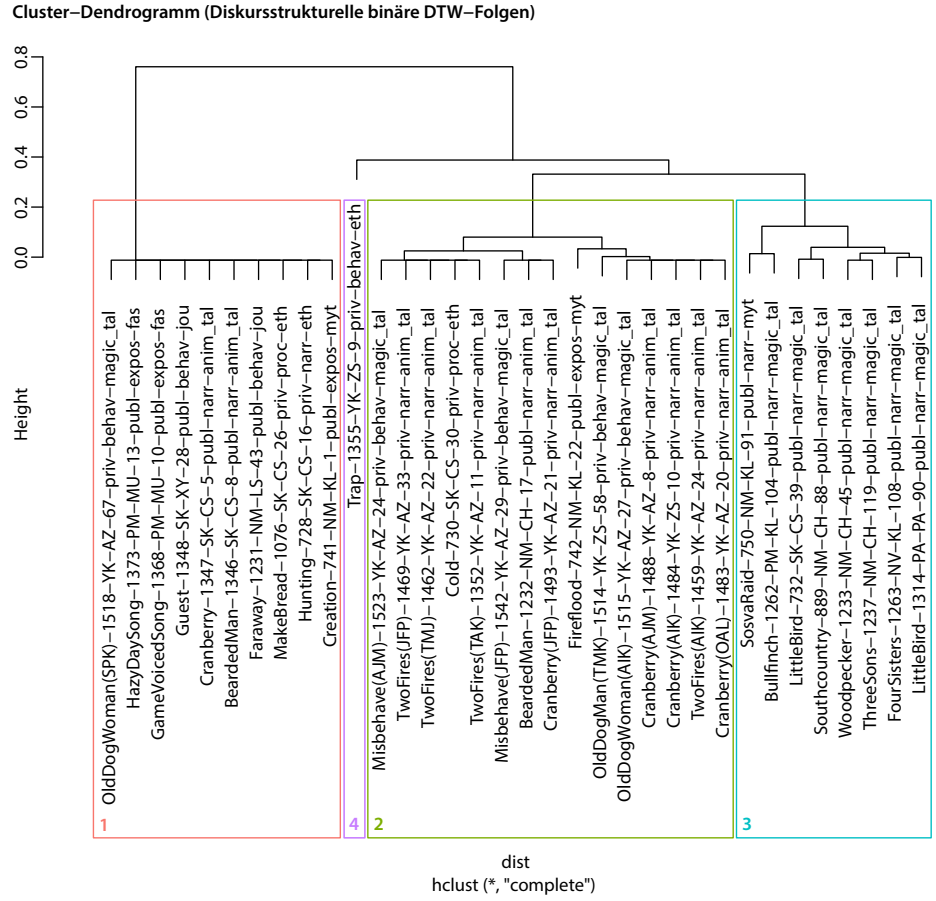


Plot P.3: Sequenz-Indexplot für GENRE-Klassifizierung (Diskursstrukturelle Sequenzen)

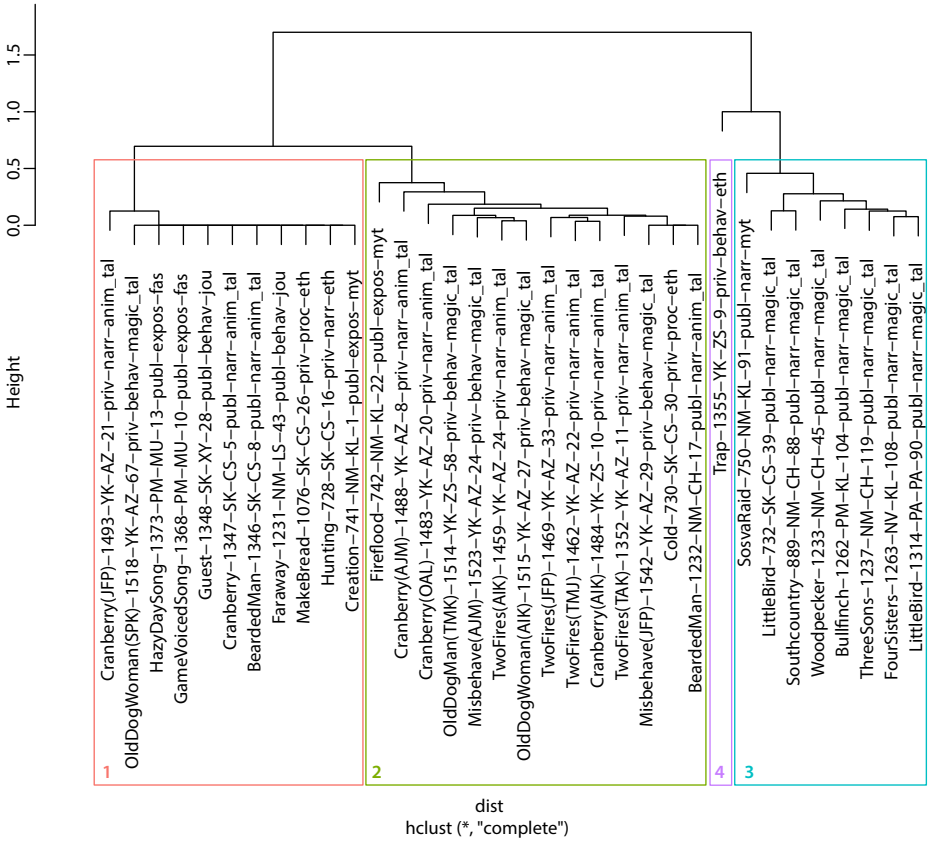


Plot P.4: Durchschnittliche Verweildauer in GENRE-Klassen (Diskursstrukturelle Sequenzen)

Q Plots zu 6.5.7 (Diskursstrukturelle Partitur-Folgen)

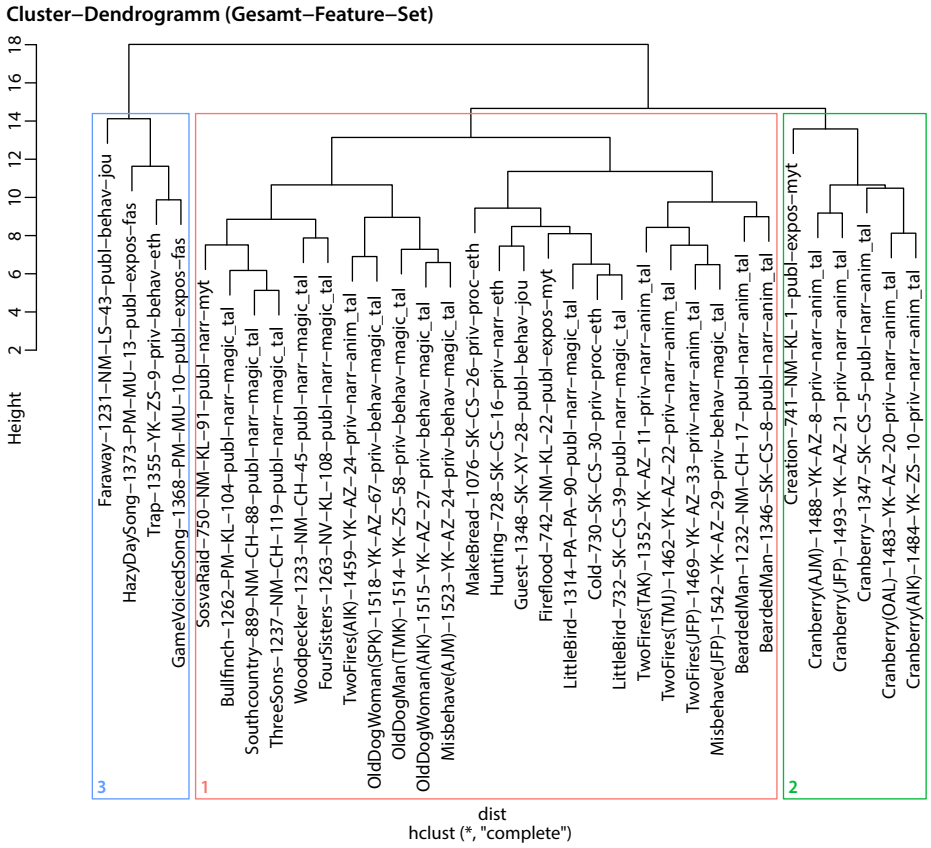


Plot Q.1: Cluster-Dendrogramm (Binäre DTW-Diskurspartitur)

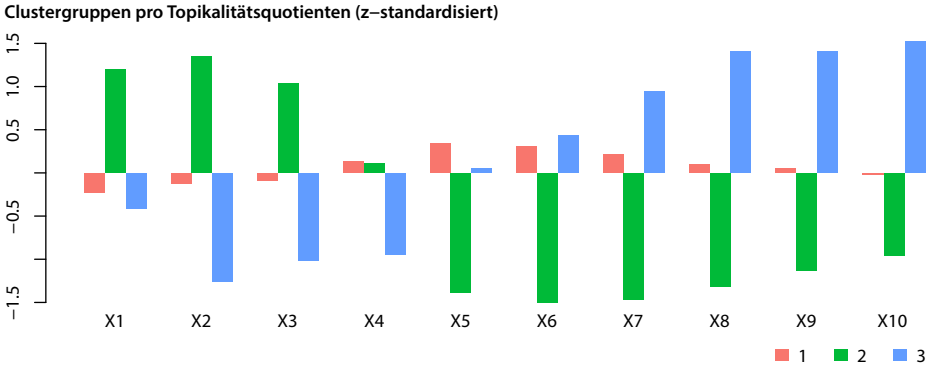


Plot Q.2: Cluster-Dendrogramm (Aggregierte DTW-Diskurspartitur)

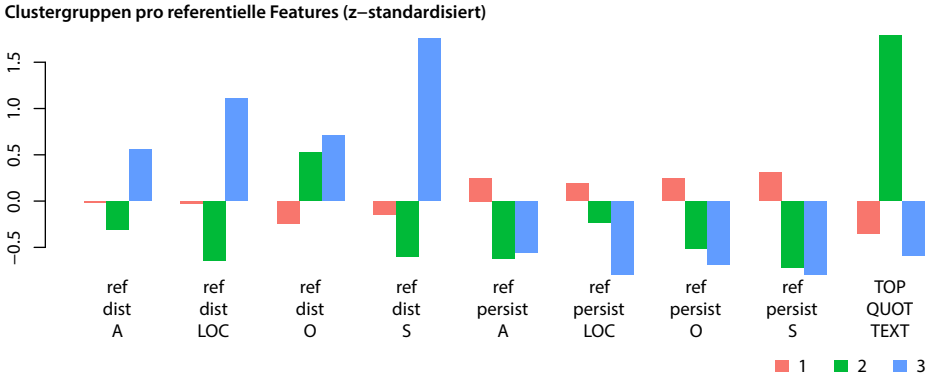
R Plots zu 6.6.1 (Feature-basiertes Gesamtmodell)



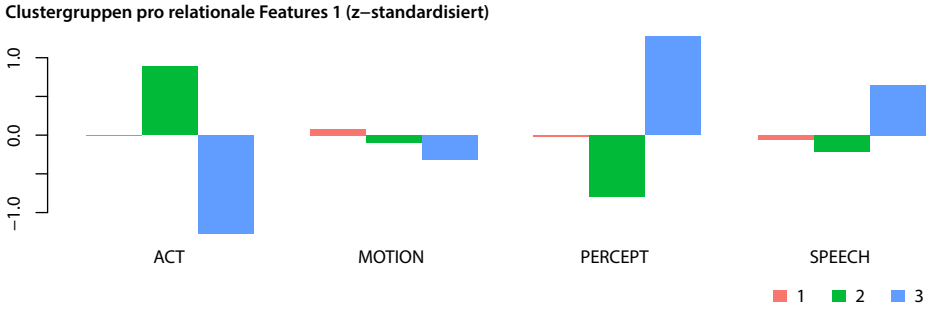
Plot R.1: Cluster-Dendrogramm (Gesamt-Feature-Set)



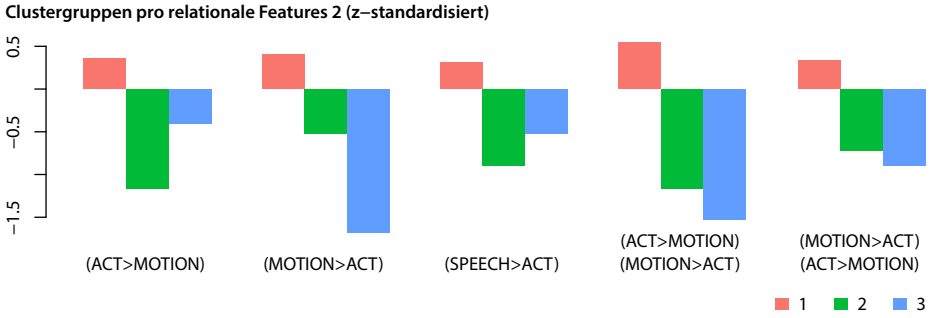
Plot R.2: Average-Scores-Barplots der Clustergruppen, gruppiert nach Topikalitätsquotienten (Gesamt-Feature-Set)



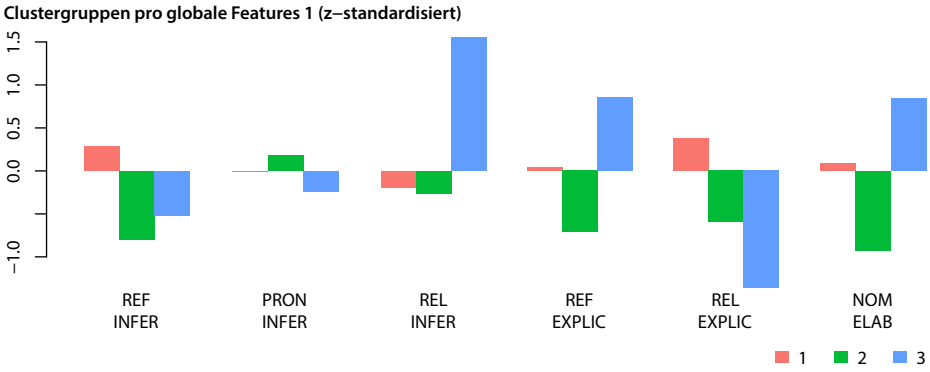
Plot R.3: Average-Scores-Barplots der Clustergruppen, gruppiert nach referentiellen Features (Gesamt-Feature-Set)



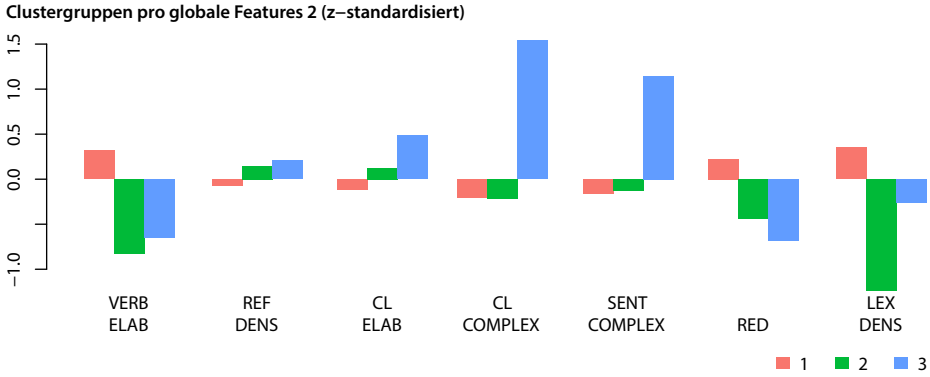
Plot R.4: Average-Scores-Barplots der Clustergruppen für Ereignistypik (Gesamt-Feature-Set)



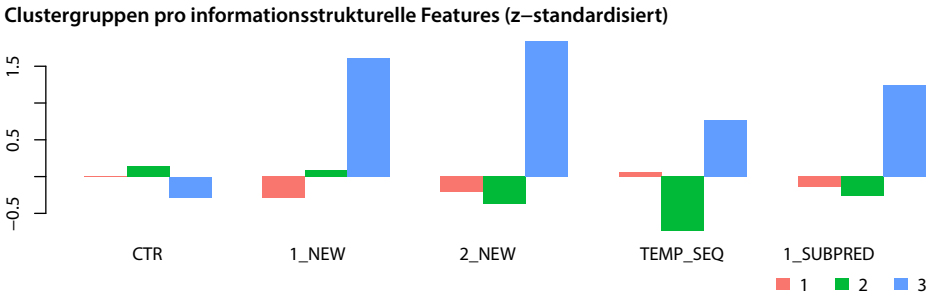
Plot R.5: Average-Scores-Barplots der Clustergruppen für Ereignisübergänge (Gesamt-Feature-Set)



Plot R.6: Average-Scores-Barplots der Clustergruppen, gruppiert nach globalen Features 1 (Gesamt-Feature-Set)



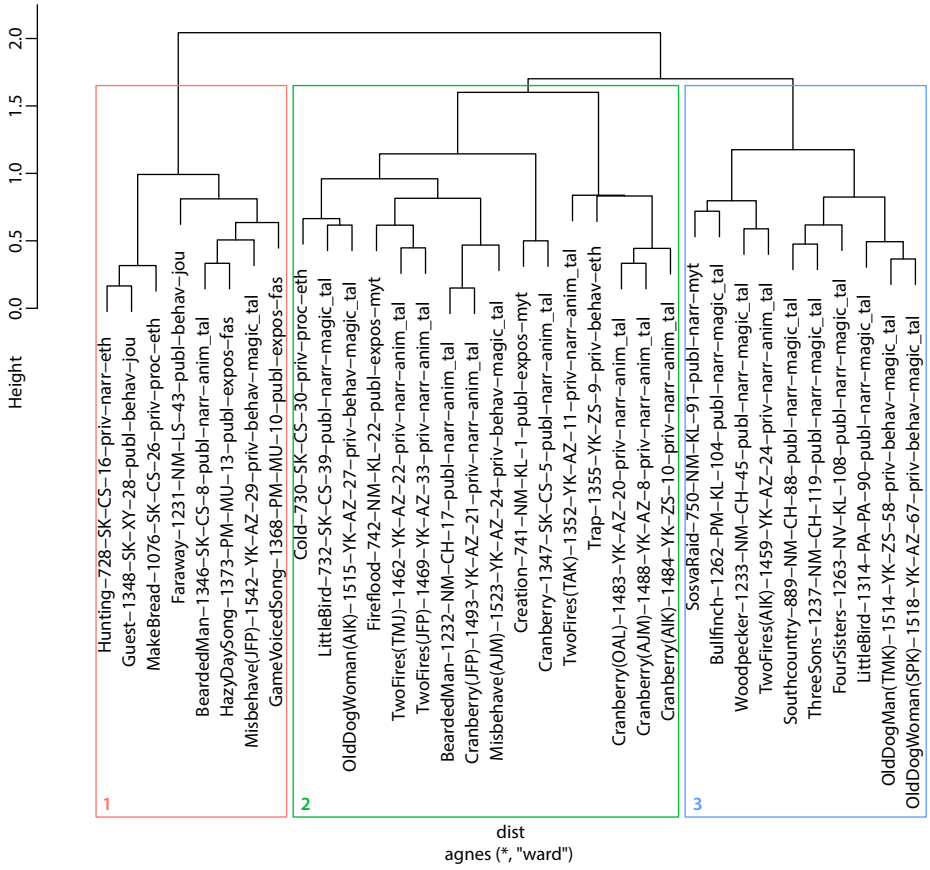
Plot R.7: Average-Scores-Barplots der Clustergruppen, gruppiert nach globalen Features 2 (Gesamt-Feature-Set)



Plot R.8: Average-Scores-Barplots der Clustergruppen, gruppiert nach informationsstrukturellen Features (Gesamt-Feature-Set)

S Plots zu 6.7.2 (Diskussion der Sequenzanalysen)

Cluster–Dendrogramm (Switch–Reference–Folgen, smoothed)



Plot S.1: Cluster-Dendrogramm (Switch-Reference-Sequenzen mit Smoothing)

Abbildungsverzeichnis

Tabellenverzeichnis

Plotverzeichnis

Reportverzeichnis

Auflistungsverzeichnis

Annotationsverzeichnis

Die folgenden Listen und Beschreibungen sind aus der OUDB-Datenbank extrahiert sowie den Dokumentationen des OUDB-Korpus entnommen.

Part-of-Speech-Tags

adj:Any
adj>adj
adj>subs
adj
adv:Any
adv
appnum
cardnum:Any
cardnum>ordnum
cardnum>reitnum
cardnum>subs
cardnum
cconj
cop
dem.dist
dem.prox
dem
det
epers
epro
fil
IDPH
indfpro
interj
interrog
ipro
irogpro
n:Any
n>adj
n>n
n>v
neg.exist
neg.indf.pron
neg.pron
neg.ptcl
negexist
nfin:Any
nprop
num:Any
num
n
ordnum
particle
ppron
pro-form:Any
pro-form
pro:Any
pro>adj
pro>subs
pro
prvb
pstp:Any
pstp
ptcl
ptcp:Any
ptcp>subs
quant
quest. ptcl
reitnum
rhyt
rpro
sconj
spers
subs:Any
subs>adj
subs>adv
subs
v:Any
v>adj
v>cvb
v>inf
v>nfin
v>ptcp
v>v
v

Morphologische Glossen

1
2
3
<
ABE
ABL
ADJZR.CAR
ADJZR.PROP
ADJZR
ADVZR
APP
AUG
CAUS
COLL
COMPR

COMP
COM
COND
CVB
DEG
DIM
DLAT
DUR
DU
EXPL
FREQ
IMP.CLM
IMP
INCH
INF
INSC
INST
INTR
IRRFL
LOC
MEL
MIR
MOM
NEG
NOMZR
NON-SG
NZER
OBL
OPT
ORD
PASS
PEJ
PFV
PL
PREC
PROPR
PRS
PST
PTCP
REFL
REITNUM
RES
SG
TRANS
TRNS
VBZR
VOC
VZER

Annotationsparameter

Syntaktische Annotationsparameter

adjP Adjective phrase
advP Adverbial phrase
attrP Attributive phrase
compC Complement clause
CONJ Conjunction
CVB Converb
finVP Finite verbal phrase
infVP Infinite (verbal) phrase
locNP Noun phrase with local case
NEG Negation
NP Noun phrase
NUM Numeral
okVP Finite verbal phrase with objective conjugation
passVP Finite verbal phrase in passive voice
postP Postpositional phrase
pronP Pronominal phrase
PTCL Particle
ptcpVP Participial (verbal) phrase
px Possessive suffix
QP Interrogative
subC Subordinated clause
zero Zero anaphora

Funktionale Annotationsparameter

ADV	Adverbial (Place, Time, etc.)
AGR	Agreement
ATTR	Attribute preceding the head of a phrase
COLL	Collocation
CON	Connector
IO	Indirect Object of a sentence
MOD	Modifier of a phrase (cf. possessive suffixes)
O	Direct Object of a sentence
PAR	Parenthesis, insertion with no syntactic function
PRED	Predicate of a sentence
S	Subject of a sentence
SUBPRED	Predicate of a subordinate clause
VOC	Vocative, constituent of a sentence (i. e. a referent) with no other syntactic role other than being the addressee of the utterance

Semantische Rollen

ADD	Additive
ADR	Addressee (animate)
ADVERS	Adversative
AG	Agent
CAUSE	Cause
COM	Comitative (animate)
COMP	Comparative
CONS	Consecutive
DEG	Degree
DISJ	Disjunctive
GOAL	Goal
INST	Instrument (inanimate)
LOC	Location
MANNER	Manner
PAT	Patient
PATH	Path
PWH	Part-Whole Relation (for possessive suffixes as marker of collocations)
QUAL	Quality
REC	Recipient
REF	Reference-Point
SOURCE	Source
TIME	Time

Semantische Klassen

Nominale Klassen

ANIM	Animate
BODY	Body part
HUM	Human
INANIM	Inanimate

Verbale Klassen

ACT	Action&Process
MOTION	Motion
PERCEPT	Perception
SPEECH	Speech
STATE	State

Pragmatische Annotationsparameter

ANCH	Anchoring
CTR	kontrastiver Fokus
DISC	Diskursmarker
FOC	Fokus
FRAME	Frame-Setting
KG	Known Group
MFOC	Mirror Focus
REPEAT	unmittelbar wiederholter Referent/Handlung
RESUME	wiederaufgenommener Referent
THL	Tail Head Linkage
TOP	Topik

Abkürzungsverzeichnis

Allgemeine Abkürzungen

A	Agentive (grammatische Relation)
ASCII	American Standard Code for Information Interchange
ASW	Average Silhouette Width
ATU	Aarne-Thompson-Uther-Index
avscores	Average Scores Barplot
Bagging	Bootstrap Aggregation
clusplot	PCA-Clusterplot
complete	Complete-Linkage-Agglomerationsmethode
coordpar	Parallelkoordinatenplot
cv	Kreuzvalidierung
DTW	Dynamic Time Warping
euclid	Euklidisches Distanzmaß
ev	Ereignisvorstellung
evs	Ereignisstruktur-bezogene Modelle
feat_import	Feature-Importance
indel	Insertion/Deletion
INFO-SPEECH	textinterner diskursiver Status
IO	Indirect Objective (grammatische Relation)
knn	k-Nearest-Neighbour
LDA	Lineare Diskriminanzanalyse
LM	Landmark
LOC	Locative (grammatische Relation)
ML	Machine Learning
mtry	Anzahl der in RF pro Split verwendeten Merkmale
NA	not available
NLP	Natural Language Processing
O	Objective (grammatische Relation)
OM	Optimal Matching
OUDB	Ob-Ugric Database
PCA	Hauptkomponentenanalyse
POS	Part-of-Speech
px	Possessivsuffix
ref	referentiell
refs	Referentenstruktur-bezogene Modelle
rel	relational
RF	Random Forest
S	Subjective (grammatische Relation)
seq	Sequenzfolge
single	Single-Linkage-Agglomerationsmethode
SQL	Structured Query Language
subj. Konj	subjektive Konjugation
obj. Konj	objektive Konjugation
SVM	Support Vector Machine
TAM	Tempus-Aspekt-Modus
trans	Übergang (Transition)
TWM	Text-Weltmodell
USAS	UCREL Semantic Analysis System
ward	Wards Agglomerationsmethode
WSS	Mittel der Fehlerquadrate
Ø	Nullmorphem (zero)

Feature-Abkürzungen

Für die Feature-Werte der Ereignistypik (**ACT**, **MOTION**, **PERCEPT**, **SPEECH**) sowie die daraus zusammengesetzten Feature-Werte der Ereignisabfolge s. das Verzeichnis der verbalen semantischen Klassen der Annotationsparameter; für die Feature-Abkürzungen der pragmatischen Typik s. das Verzeichnis der pragmatischen Annotationsparameter; für die Feature-Werte der Subordinationstypen s. das Verzeichnis der syntaktischen Annotationsparameter.

1/2_NEW	Topik-Einführung in erster/zweiter Texthälfte
1/2_SUBPRED	Subordinationsstärke in erster/zweiter Texthälfte
CL_COMPLEX	Clause-Komplexität
CL_ELAB	Clause-Elaboration
LEX_DENS	lexikalische Dichte
NEW	Topik-Einführung (Informationsdichte)
NOM_ELAB	nominale Elaboration
PRON_INFER	pronominaler Inferenzgrad
RED	Redundanz
REF_DENS	referentielle Dichte
REF_EXPLIC	referentielle Explizitheit
REF_INFER	referentieller Inferenzgrad
ref-dist	referentielle Distanz
ref-persist	Topik-Persistenz
REL_EXPLIC	relationale Explizitheit
REL_INFER	relationaler Inferenzgrad
SENT_COMPLEX	Satz-Komplexität
SUBORD	Subordinationstypen
SUBPRED	Subordinationsstärke
TEMP_SEQ	Temporal-Sequencing
TOP_QUOT	durchschnittlicher Topikalitätsquotient
TOP_QUOT_TEXT	textweiter Topikalitätsquotient
VERB_ELAB	verbale Elaboration
X1, X2, ..., X10	Topikalitätsquotient des häufigsten, zweithäufigsten, ..., zehnthäufigsten Referenten

Sequentielle Feature-Abkürzungen

CONT	same subject	**SPEECH**	eingebetteter Dialog
NONE	keine Einbettung	**SWITCH**	switch subject

Abkürzungen der Textklassifizierungen

Textklassifizierungen

BASE	Basis-Textsorten-Klassifizierung
BINARY	binäre Kategorisierung (narrativ vs. nicht-narrativ)
COMM_SIT	Textklassifizierung nach Kommunikationstyp
DISC_STRUCT	diskursstrukturelle Textklassifizierung
GENRE	Subklassifizierung der Basis-Textsorten nach narrativen Subgenres (Zaubermärchen vs. Tiermärchen)

Abkürzungen der Textklassen

anim_tal	Fabel-Tiermärchen	**myt**	mythologische Sagen
behav	verhaltensbezogener Diskurs	**narr**	narrativer Diskurs
eth	ethnographische Texte	**priv**	Interview (privat)
expos	expositorischer Diskurs	**proc**	prozeduraler Diskurs
fas	Personal Songs (Fate Songs)	**publ**	öffentlicher Vortrag
jou	journalistische Berichte	**tal**	Volksmärchen (Tales)
magic_tal	Zaubermärchen		

Abkürzungen der Sammlungen und Dialekte

Kürzel der Sammlungen

AZ	Kayukova & Schön	**LS**	Zeitung *Luima Seripos*
CH	Chernetsov, Valeri	**PA**	Paasonen, Heikki
CS	Csepregi, Márta	**XY**	Zeitung *Khanty Yasang*
KL	Kannisto & Liimola	**ZS**	Schön, Zsófia

Dialektkürzel

NM Nordmansi
NV Nordvagilsk Mansi
PA Yugan Khanty (Paasonen Korpus)
PM Pelym Mansi
SK Surgut Khanty
YK Yugan Khanty

Korpusverzeichnis

728 [Hunting]: Hunting Adventure. Csepregi, Márta 1998: OUDB Surgut Khanty Corpus. Text ID 728. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=728 [abgerufen am 18.03.2019]

730 [Cold]: An Old Story. Csepregi, Márta 1998: OUDB Surgut Khanty Corpus. Text ID 730. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=730 [abgerufen am 18.03.2019]

732 [LittleBird]: The Little Bird and His Sister. Csepregi, Márta 1998: OUDB Surgut Khanty Corpus. Text ID 732. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=732 [abgerufen am 18.03.2019]

741 [Creation]: Creation of the Earth. Kannisto & Liimola 1951: OUDB Northern Mansi Corpus. Text ID 741. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=741 [abgerufen am 18.03.2019]

742 [Fireflood]: The Holy Fireflood. Kannisto & Liimola 1951: OUDB Northern Mansi Corpus. Text ID 742. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=742 [abgerufen am 18.03.2019]

750 [SosvaRaid]: The Middle Sosva Old Man's Raid to the Sacred Site on the Water. Kannisto & Liimola 1951: OUDB Northern Mansi Corpus. Text ID 750. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=750 [abgerufen am 18.03.2019]

889 [Southcountry]: Southcountry. Chernetsov, Valeri 1990: OUDB Northern Mansi Corpus. Text ID 889. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=889 [abgerufen am 18.03.2019]

1076 [MakeBread]: Make Bread. Csepregi, Márta 2002: OUDB Surgut Khanty Corpus. Text ID 1076. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1076 [abgerufen am 18.03.2019]

1231 [Faraway]: We Studied in a Faraway Country. Luima Seripos 2005: OUDB Northern Mansi Corpus. Text ID 1231. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=1231 [abgerufen am 18.03.2019]

1232 [BeardedMan]: A Tale of Four Men. Chernetsov, Valeri: OUDB Northern Mansi Corpus. Text ID 1232. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=1232 [abgerufen am 18.03.2019]

1233 [Woodpecker]: The Poor Old Woman, Her Husband and the Woodpecker. Chernetsov, Valeri: OUDB Northern Mansi Corpus. Text ID 1233. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=1233 [abgerufen am 18.03.2019]

1237 [ThreeSons]: In Olden Times There Lived a Man Who Had Three Sons. Chernetsov, Valeri 1933: OUDB Northern Mansi Corpus. Text ID 1237. Hrsg. von Janda, Gwen Eva. http://www.oudb.gwi.uni-muenchen.de/?cit=1237 [abgerufen am 18.03.2019]

1262 [Bullfinch]: There Was an Old Man and an Old Woman. Kannisto & Liimola 1956: OUDB Pelym Mansi Corpus. Text ID 1262. Hrsg. von Eichinger, Viktória. http://www.oudb.gwi.uni-muenchen.de/?cit=1262 [abgerufen am 18.03.2019]

1263 [FourSisters]: Four Sisters and a Man with His Daugther. Kannisto & Liimola 1956: OUDB Northern Vagilsk Mansi Corpus. Text ID 1263. Hrsg. von Wolfauer, Anna. http://www.oudb.gwi.uni-muenchen.de/?cit=1263 [abgerufen am 18.03.2019]

1314 [LittleBird]: The Little Bird and His Sister. Paasonen, Heikki 2001: OUDB Yugan Khanty (1901) Corpus. Text ID 1314. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1314 [abgerufen am 18.03.2019]

1346 [BeardedMan]: Two Short Stories 1. Csepregi, Márta 2011: OUDB Surgut Khanty Corpus. Text ID 1346. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1346 [abgerufen am 18.03.2019]

1347 [Cranberry]: Two Short Stories 2. Csepregi, Márta 2011: OUDB Surgut Khanty Corpus. Text ID 1347. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1347 [abgerufen am 18.03.2019]

1348 [Guest]: Our Guest. Khanty Yasang 2012: OUDB Surgut Khanty Corpus. Text ID 1348. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1348 [abgerufen am 18.03.2019]

1352 [TwoFires]: Two Domestic Fires (TAK). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1352. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1352 [abgerufen am 18.03.2019]

1355 [Trap]: Trap. Schön, Zsófia 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1355. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1355 [abgerufen am 18.03.2019]

1368 [GameVoicedSong]: Song Composed by Agsinia Apanasejevna. Munkácsi, Bernát 1896: OUDB Pelym Mansi Corpus. Text ID 1368. Hrsg. von Eichinger, Viktória. http://www.oudb.gwi.uni-muenchen.de/?cit=1368 [abgerufen am 18.03.2019]

1373 [HazyDaySong]: Song Composed by Agafia Stefanovna. Munkácsi, Bernát 1896: OUDB Pelym Mansi Corpus. Text ID 1373. Hrsg. von Eichinger, Viktória. http://www.oudb.gwi.uni-muenchen.de/?cit=1373 [abgerufen am 18.03.2019]

1459 [TwoFires]: Two Domestic Fires (AIK). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1459. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1459 [abgerufen am 18.03.2019]

1462 [TwoFires]: Two Domestic Fires (TMJ). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1462. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1462 [abgerufen am 18.03.2019]

1469 [TwoFires]: Two Domestic Fires (JFP). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1469. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1469 [abgerufen am 18.03.2019]

1483 [Cranberry]: Little Cranberry (OAL). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1483. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1483 [abgerufen am 18.03.2019]

1484 [Cranberry]: Little Cranberry (AIK). Schön, Zsófia 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1484. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1484 [abgerufen am 18.03.2019]

1488 [Cranberry]: Little Cranberry (AJM). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1488. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1488 [abgerufen am 18.03.2019]

1493 [Cranberry]: Little Cranberry (JFP). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1493. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1493 [abgerufen am 18.03.2019]

1514 [OldDogMan]: Old-Dog-Backtendoned-Man (TMK). Schön, Zsófia 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1514. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1514 [abgerufen am 18.03.2019]

1515 [OldDogWoman]: Old-Dog-Backtendoned-Woman (AIK). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1515. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1515 [abgerufen am 18.03.2019]

1518 [OldDogWoman]: Old-Dog-Backtendoned-Woman (SPK). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1518. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1518 [abgerufen am 18.03.2019]

1523 [Misbehave]: Why One Shouldn't Misbehave at Night (AJM). Kayukova & Schön 2016: OUDB Yugan Khanty (2010–) Corpus. Text ID 1523. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1523 [abgerufen am 18.03.2019]

1542 [Misbehave]: Why One Shouldn't Misbehave at Night (JFP). Kayukova & Schön 2017: OUDB Yugan Khanty (2010–) Corpus. Text ID 1542. Hrsg. von Schön, Zsófia. http://www.oudb.gwi.uni-muenchen.de/?cit=1542 [abgerufen am 18.03.2019]

Literaturverzeichnis

Abbott, Andrew & Tsay, Angela (2000): Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research* 29.1, 3–33. DOI: 10.1177/0049124100029001001.

Abondolo, Daniel (1998a): Introduction. In: *The Uralic Languages*. Hrsg. von Daniel Abondolo. London & New York: Routledge, 1–42.

— (1998b): Khanty. In: *The Uralic Languages*. Hrsg. von Daniel Abondolo. London & New York: Routledge, 358–386.

Aggarwal, Charu C. (2015): *Data Mining*. Cham: Springer. DOI: 10.1007/978-3-319-14142-8.

— (2018): *Machine Learning for Text*. Cham: Springer. DOI: 10.1007/978-3-319-73531-3.

Altmann, Eduardo G. & Gerlach, Martin (2016): Statistical Laws in Linguistics. In: *Creativity and Universality in Language*. Hrsg. von Mirko Degli Esposti & Eduardo G. Altmann & François-David Pachet. Cham: Springer, 7–26. DOI: 10.1007/978-3-319-24403-7_2.

Andersen, Flemming G. (1997): Rezension: Rubin, David C. Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes. 1995. *Jahrbuch für Volksliedforschung* 42, 173–176. DOI: 10.2307/848046.

Anderson, John R. (1975): Computer Simulation of a Language Acquisition System: A First Report. In: *Information Processing and Cognition: The Loyola Symposium*. Hrsg. von Robert L. Solso. Washington: Erlbaum, 295–349. DOI: 10.1184/R1/6614087.v1.

Archer, Dawn & Wilson, Andrew & Rayson, Paul (2002): *Introduction to the USAS Category System*. URL: http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf [abgerufen am 14. 07. 2021].

Augoustinos, Martha & Walker, Iain & Donaghue, Ngaire (2006): *Social Cognition: An Integrated Introduction*. London: SAGE. URI: https://psycnet.apa.org/record/2006-08658-000.

Azzalini, Adelchi & Menardi, Giovanna (2014): Clustering via Nonparametric Density Estimation: The R Package pdfCluster. *Journal of Statistical Software* 57.11, 1–26. DOI: 10.18637/jss.v057.i11.

Banfield, Ann (1973): Narrative Style and the Grammar of Direct and Indirect Speech. *Foundations of Language* 10.1, 1–39. URI: https://www.jstor.org/stable/25000702.

Bartlett, Frederic C. (1932): *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press. URI: https://psycnet.apa.org/record/1995-98505-000.

Bascom, William (1965): The Forms of Folklore: Prose Narratives. *The Journal of American Folklore* 78.307, 3–20. DOI: 10.2307/538099.

Bawarshi, Anis S. & Reiff, Mary Jo (2010): *Genre: An Introduction to History, Theory, Research, and Pedagogy*. West Lafayette, IN: Parlor Press. URI: https://wac.colostate.edu/books/referenceguides/bawarshi-reiff.

Beneš, Eduard (1973): Thema-Rhema-Gliederung und Textlinguistik. In: *Studien zur Texttheorie und zur deutschen Grammatik. Festgabe für Hans Glinz zum 60. Geburtstag*. Hrsg. von Horst Sitta & Klaus Brinker. Sprache der Gegenwart 30. Düsseldorf: Schwann, 42–62. URI: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-22407.

Beyerer, Jürgen & Richter, Matthias & Nagel, Matthias (2017): *Pattern Recognition. Introduction, Features, Classifiers and Principles*. Berlin & Boston: De Gruyter. DOI: 10.1515/9783110537949.

Biber, Douglas (1992a): The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26.5-6, 331–345. DOI: 10.1007/BF00136979.

— (1992b): Using Computer-Based Text Corpora to Analyze the Referential Strategies of Spoken and Written Texts. In: *Directions in Corpus Linguistics*. Hrsg. von Jan Svartvik. Trends in Linguistics. Studies and Monographs 65. Berlin & New York: De Gruyter, 213–252. DOI: 10.1515/9783110867275.213.

— (1995): *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511519871.

— (2014): Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation. *Languages in Contrast* 14.1, 7–34. DOI: 10.1075/lic.14.1.02bib.

Biber, Douglas & Conrad, Susan (2001): Register Variation: A Corpus Approach. In: *The Handbook of Discourse Analysis*. Hrsg. von Deborah Schiffrin & Deborah Tannen & Heidi E. Hamilton. Malden, MA: Blackwell, 175–196. DOI: 10.1002/9780470753460.ch10.

Biber, Douglas & Conrad, Susan & Reppen, Randi (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Approaches to Linguistics. Cambridge [u. a.]: Cambridge University Press. DOI: 10.1017/CBO9780511804489.

Biber, Douglas & Jones, James K. (2009): Quantitative Methods in Corpus Linguistics. In: *Corpus Linguistics*. Hrsg. von Anke Lüdeling & Merja Kytö. Bd. 2. Handbücher zur Sprach- und Kommunikationswissenschaft 29. Berlin & New York: De Gruyter, 1286–1304. DOI: 10.1515/9783110213881.2.1286.

Bickel, Balthasar (2003): Referential Density in Discourse and Syntactic Typology. *Language* 79.4, 708–736. DOI: 10.1353/lan.2003.0205.

Birnbaum, Marianna D. (1977): Rezension: Dégh, Linda. People in the Tobacco Belt: Four Lives. 1975. *East Central Europe* 4, 231.

Bossong, Georg (1998): Le marquage différentiel de l'objet dans les langues d'Europe. In: *Actance et valence dans les langues de l'Europe*. Hrsg. von Jack Feuillet. Bd. 2. Empirical Approaches to Language Typology 20. Berlin: De Gruyter, 193–258. DOI: 10.1515/9783110804485.193.

Bredenkamp, Jürgen (2020): Gedächtnis. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 660f.

Breiman, Leo (1996): Bagging Predictors. *Machine Learning* 24.2, 123–140. DOI: 10.1007/BF00058655.

— (2001): Random Forests. *Machine Learning* 45.1, 5–32. DOI: 10.1023/A:1010933404324.

— (2002): *Manual On Setting Up, Using, And Understanding Random Forests V3.1*. URL: https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf [abgerufen am 21.05.2020].

Breiman, Leo & Cutler, Adele (2020): *Random Forests*. Dokumentation. URL: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [abgerufen am 19.05.2020].

Brewer, William F. & Nakamura, Glenn V. (1984): *The Nature and Functions of Schemas*. Center for the Study of Reading, Technical Report 325. Champaign, IL: University of Illinois at Urbana-Champaign. URI: http://hdl.handle.net/2142/17542.

Brillinger, David R. (2011): Data Analysis, Exploratory. In: *International Encyclopedia of Political Science*. Hrsg. von Bertrand Badie & Dirk Berg-Schlosser & Leonardo Morlino. Thousand Oaks, CA: SAGE, 531–538. DOI: 10.4135/9781412959636.n128.

Brinker, Klaus (2000): Textfunktionale Analyse. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 175–186. DOI: 10.1515/9783110194067.

— (2018): *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. Hrsg. von Hermann Cölfen & Steffen Pappert. Berlin: Erich Schmidt.

Brüder Grimm (2010): *Kinder- und Hausmärchen*. Stuttgart: Reclam.

Bubenhofer, Noah (2008): Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse. In: *Methoden der Diskurslinguistik: Sprachwissenschaftliche Zugänge zur transtextuellen Ebene*. Hrsg. von Jürgen Spitzmüller & Ingo H. Warnke. Berlin & New York: De Gruyter, 407–434. DOI: 10.1515/9783110209372.6.407.

Bußmann, Hadumod (2008): *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner.

Carletta, Jean (1996): Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22.2, 249–254. URI: https://aclanthology.org/J96-2004.

Chafe, Wallace L. (1976): Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: *Subject and Topic*. Hrsg. von Charles N. Li. New York: Academic Press, 25–56.

— (1987): Cognitive Constraints on Information Flow. In: *Coherence and Grounding in Discourse*. Hrsg. von Russell S. Tomlin. Typological Studies in Language 11. Amsterdam [u. a.]: Benjamins, 21–51. DOI: 10.1075/tsl.11.03cha.

— (2001): The Analysis of Discourse Flow. In: *The Handbook of Discourse Analysis*. Hrsg. von Deborah Schiffrin & Deborah Tannen & Heidi E. Hamilton. Malden, MA: Blackwell, 673–687. DOI: 10.1002/9780470753460.ch35.

Chambers, Nathanael & Jurafsky, Dan (2009): Unsupervised Learning of Narrative Schemas and Their Participants. 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*. Bd. 2. ACL '09. Singapur: Association for Computational Linguistics, 602–610. DOI: 10.3115/1690219.1690231.

Chernetsov, Valeri N. (1963): Concepts of the Soul among the Ob Ugrians. In: *Studies in Siberian Shamanism No. 4*. Hrsg. von Henry Michael. Toronto: University of Toronto Press, 1–45. DOI: 10.3138/9781487589509-002.

Cooreman, Ann M. (1987): *Transitivity and Discourse Continuity in Chamorro Narratives*. Empirical Approaches to Language Typology 4. Berlin & New York: De Gruyter. DOI: 10.1515/9783110851014.

Croft, William (2016): Typology and the Future of Cognitive Linguistics. *Cognitive Linguistics* 27.4, 587–602. DOI: 10.1515/cog-2016-0056.

Csepregi, Márta (2005): Das Vöglein und seine Schwester – Variationen eines obugrischen Märchentyps. In: *Lihkkun lehkos! Beiträge zur Finnougristik aus Anlaß des sechzigsten Geburtstages von Hans-Hermann Bartens*. Hrsg. von Cornelius Hasselblatt & Eino Koponen & Anna Widmer. Veröffentlichungen der Societas Uralo-Altaica 65. Wiesbaden: Harrassowitz, 57–64.

— (2009): The Very Highly Connected Nodes in the Ob-Ugrian Networks. In: *The Quasquicentennial of the Finno-Ugrian Society*. Hrsg. von Jussi Ylikoski. Suomalais-Ugrilaisen Seuran Toimituksia. Mémoires de la Société Finno-Ougrienne 258, 9–32. URI: https://www.sgr.fi/en/items/show/683.

Cumming, Susanna & Ono, Tsuyoshi & Laury, Ritva (2011): Discourse, Grammar and Interaction. In: *Discourse Studies: A Multidisciplinary Introduction*. Hrsg. von Teun A. van Dijk. London: SAGE, 8–36. DOI: 10.4135/9781446289068.n2.

Cushing, George F. (1980): Ob Ugrian (Vogul and Ostyak). In: *Traditions of Heroic and Epic Poetry*. Hrsg. von Arthur T. Hatto & Robert Auty. Bd. 1. London: Modern Humanities Research Association, 211–235.

Daller, Michael (2010): Guirauds Index of Lexical Richness. In: *Conference of the British Association of Applied Linguistics*. URI: http://eprints.uwe.ac.uk/11902.

Daneš, František (1970): Zur linguistischen Analyse der Textstruktur. *Folia Linguistica* 4.1, 72–78. DOI: 10.1515/flin.1970.4.1-2.72.

Day, Peter (2020): Raum. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 1478.

De Beaugrande, Robert-Alain & Dressler, Wolfgang U. (1981): *Einführung in die Textlinguistik*. Konzepte der Sprach- und Literaturwissenschaft 28. Tübingen: Niemeyer. DOI: 10.1515/9783111349305.

DeLamater, John D. & Myers, Daniel J. & Collett, Jessica L. (2018): *Social Psychology*. Boulder: Routledge. DOI: 10.4324/9780429493096.

Deza, Michel M. & Deza, Elena (2016): *Encyclopedia of Distances*. Berlin & Heidelberg: Springer. DOI: 10.1007/978-3-662-52844-0.

Dik, Simon C. (1991): Functional Grammar. In: *Linguistic Theory and Grammatical Description: Nine Current Approaches*. Hrsg. von Flip G. Droste & John E. Joseph. Current Issues in Linguistic Theory 75. Amsterdam: Benjamins, 247–274. DOI: 10.1075/cilt.75.09dik.

— (1997a): *The Theory of Functional Grammar. Part 1: The Structure of the Clause*. Functional Grammar Series 20. Berlin & New York: De Gruyter. DOI: 10.1515/9783110218367.

— (1997b): *The Theory of Functional Grammar. Part 2: Complex and Derived Constructions*. Functional Grammar Series 21. Berlin & New York: De Gruyter. DOI: 10.1515/9783110218374.

Ding, Hui u. a. (2008): Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. *Proceedings of the VLDB Endowment* 1.2, 1542–1552. DOI: 10.14778/1454159.1454226.

Dithmar, Reinhard (1974): *Die Fabel: Geschichte, Struktur, Didaktik*. Paderborn: Schöningh.

Dorgeloh, Heidrun & Wanner, Anja, Hrsg. (2010): *Syntactic Variation and Genre*. Topics in English Linguistics 70. Berlin & New York: De Gruyter. DOI: 10.1515/9783110226485.

Dreisbach, Gesine (2020): Embedded-Processes-Modell des Arbeitsgedächtnisses. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 480.

Du Bois, John W. (1987): The Discourse Basis of Ergativity. *Language* 63.4, 805–855. DOI: 10.2307/415719.

— (2003): Discourse and Grammar. In: *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Hrsg. von Michael Tomasello. Bd. 2. Mahwah, NJ: Erlbaum, 47–87.

Ehrlich, Susan (1987): Aspect, Foregrounding and Point of View. *Text – Interdisciplinary Journal for the Study of Discourse* 7.4, 363–376. DOI: 10.1515/text.1.1987.7.4.363.

Engelkamp, Johannes (2020): Textstruktur. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 1781.

Eroms, Hans-Werner (2000): Der Beitrag der Prager Schule zur Textlinguistik. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 36–42. DOI: 10.1515/9783110194067.

Figge, Udo (2000): Die kognitive Wende in der Textlinguistik. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 96–104. DOI: 10.1515/9783110194067.

Finlayson, Mark Alan (2016): Inferring Propp's Functions from Semantically Annotated Text. *Journal of American Folklore* 129, 55–77. DOI: 10.5406/jamerfolk.129.511.0055.

Finn, Aidan & Kushmerick, Nicholas (2006): Learning to Classify Documents According to Genre. *Journal of the American Society for Information Science and Technology* 57.11, 1506–1518. DOI: 10.1002/asi.20427.

Fisher, Ronald A. (1936): The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics* 7.2, 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

Fix, Ulla (2008): *Texte und Textsorten – sprachliche, kommunikative und kulturelle Phänomene*. Berlin: Frank & Timme.

Fleiss, Joseph L. (1971): Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin* 76.5, 378–382. DOI: 10.1037/h0031619.

Foucault, Michel (1981): *Archäologie des Wissens*. Frankfurt am Main: Suhrkamp.

— (1991): *Die Ordnung des Diskurses*. Frankfurt am Main: Fischer.

Frank, Markus (2019): *Phorische Verkettung im Deutschen. Eine exemplarische Untersuchung anhand von Diskursrelationen der kausalen Gruppe*. Linguistik – Impulse & Tendenzen 79. Berlin & Boston: De Gruyter. DOI: 10.1515/9783110621662.

Gabadinho, Alexis u. a. (2009): *Mining Sequence Data in R with the TraMineR Package: A User's Guide*. Dokumentation. URL: http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf [abgerufen am 11. 06. 2020].

Gabadinho, Alexis u. a. (2011): Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40.4, 1–37. DOI: 10.18637/jss.v040.i04.

Gaschler, Robert (2020): Computermetapher. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 366.

Gautier, Laurent (2009): Nochmals zum (Fach-)Textmuster: von der Kognition zur Beschreibung einzelner Textexemplare. *Lyon Linguistique allemande*. Université Lyon 2, Laboratoire LCE (Langues et Cultures Européennes). Histoire de Textes – Mélanges pour Marie-Hélène Pérennec, 1–7. URI: https://hal.archives-ouvertes.fr/hal-00425363.

Gavins, Joanna (2007): *Text World Theory: An Introduction*. Edinburgh: Edinburgh University Press. DOI: 10.1515/9780748629909.

Géron, Aurélien (2017): *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Beijing [u. a.]: O'Reilly.

Gigerenzer, Gerd (2020): Kognition. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 939f.

Giorgino, Toni (2009): Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software* 31.7, 1–24. DOI: 10.18637/jss.v031.i07.

Givón, Talmy (1983a): Introduction. In: *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Hrsg. von Talmy Givón. Amsterdam & Philadelphia: Benjamins, 1–42. DOI: 10.1075/tsl.3.01giv.

— Hrsg. (1983b): *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam & Philadelphia: Benjamins. DOI: 10.1075/tsl.3.

Goldberg, Adele E. (1995): *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago & London: University of Chicago Press.

Greimas, Algirdas J. (1972): Elemente einer narrativen Grammatik. In: *Strukturalismus in der Literaturwissenschaft*. Hrsg. von Heinz Blumensath. Köln: Kiepenheuer & Witsch, 47–67.

Gries, Stefan Th. (2008): *Statistik für Sprachwissenschaftler*. Studienbücher zur Linguistik 13. Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan Th. & Stefanowitsch, Anatol, Hrsg. (2007): *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin & New York: De Gruyter. DOI: 10.1515/9783110197709.

Grzybek, Peter & Kelih, Emmerich & Stadlober, Ernst (2005): Empirische Textsemiotik und quantitative Texttypologie. In: *Text & Reality*. Hrsg. von Jeff Bernard & Jurij Fikfak & Peter Grzybek. Ljubljana & Wien & Graz: ZRC, 95–120.

Guyon, Isabelle & Elisseeff, André (2003): An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3, 1157–1182. URI: https://dl.acm.org/doi/10.5555/944919.944968.

— (2006): An Introduction to Feature Extraction. In: *Feature Extraction. Foundations and Applications*. Hrsg. von Isabelle Guyon u. a. Studies in Fuzziness and Soft Computing 207. Berlin [u. a.]: Springer, 1–25. DOI: 10.1007/978-3-540-35488-8_1.

Guyon, Isabelle u. a. Hrsg. (2006): *Feature Extraction. Foundations and Applications*. Studies in Fuzziness and Soft Computing 207. Berlin [u. a.]: Springer. DOI: 10.1007/978-3-540-35488-8.

Häcker, Hartmut O. (2020): Internes (inneres) Modell. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus Antonius Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 863.

Han, Jiawei u. a. (2007): Frequent Pattern Mining: Current Status and Future Directions. *Data Mining and Knowledge Discovery* 15.1, 55–86. DOI: 10.1007/s10618-006-0059-1.

Harris, Zellig S. (1952): Discourse Analysis. *Language* 28.1, 1–30. DOI: 10.2307/409987.

— (1954): Distributional Structure. *WORD* 10.2-3, 146–162. DOI: 10.1080/00437956.1954.11659520.

Hartung, Wolfdietrich (2000): Kommunikationsorientierte und handlungstheoretisch ausgerichtete Ansätze. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 83–95. DOI: 10.1515/9783110194067.

Hastie, Trevor & Tibshirani, Robert & Friedman, Jerome (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer. DOI: 10.1007/978-0-387-84858-7.

Hatto, Arthur T. (2017a): Background: The Khanty. In: *The World of the Khanty Epic Hero-Princes: An Exploration of a Siberian Oral Tradition*. Cambridge: Cambridge University Press, 1–20. DOI: 10.1017/9781316216040.002.

— (2017b): *The World of the Khanty Epic Hero-Princes: An Exploration of a Siberian Oral Tradition*. Cambridge: Cambridge University Press. DOI: 10.1017/9781316216040.

Heinemann, Wolfgang (2000): Vertextungsmuster Deskription. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 356–369. DOI: 10.1515/9783110194067.

Heinemann, Wolfgang & Viehweger, Dieter (1991): *Textlinguistik*. Reihe Germanistische Linguistik 115. Tübingen: Niemeyer. DOI: 10.1515/9783111376387.

Heyer, Gerhard & Quasthoff, Uwe & Wittig, Thomas (2008): *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*. Herdecke [u. a.]: W3L.

Hofmann, Thomas & Schölkopf, Bernhard & Smola, Alexander J. (2008): Kernel Methods in Machine Learning. *The Annals of Statistics* 36.3, 1171–1220. DOI: 10.1214/009053607000000677.

Honti, László (1998): ObUgrian. In: *The Uralic Languages*. Hrsg. von Daniel Abondolo. London & New York: Routledge, 327–357.

Hopper, Paul J. (1979): Aspect and Foregrounding in Discourse. In: *Discourse and Syntax*. Hrsg. von Talmy Givón. Syntax and Semantics 12. New York [u. a.]: Brill, 213–241. DOI: 10.1163/9789004368897_010.

Hopper, Paul J. & Thompson, Sandra A. (1980): Transitivity in Grammar and Discourse. *Language* 56.2, 251–299. DOI: 10.2307/413757.

Isenberg, Horst (1978): Probleme der Texttypologie. Variation und Determination von Texttypen. *Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig* 27.5, 565–579.

— (1983): Grundfragen der Texttypologie. In: *Ebenen der Textstruktur*. Hrsg. von František Daneš & Dieter Viehweger. Linguistische Studien 112. Berlin: Akademie der Wisenschaften der DDR. Zentralinstitut für Sprachwissenschaft, 303–342.

Ismaili, Oumaima A. & Lemaire, Vincent & Cornuéjols, Antoine (2014): A Supervised Methodology to Measure the Variables Contribution to a Clustering. In: *International Conference on Neural Information Processing*. ICONIP 2014. Hrsg. von Chu K. Loo u. a. Lecture Notes in Computer Science 8834. Cham: Springer, 159–166. DOI: 10.1007/978-3-319-12637-1_20.

Jain, Anil K. & Murty, M. Narasimha & Flynn, P. J. (1999): Data Clustering: A Review. *ACM Computing Surveys* 31.3, 264–323. DOI: 10.1145/331499.331504.

James, Gareth u. a. (2017): *An Introduction to Statistical Learning: With Applications in R*. New York: Springer. DOI: 10.1007/978-1-4614-7138-7.

Janda, Gwen Eva (2015): Northern Mansi Possessive Suffixes in Non-Possessive Function. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 6.2, 243–258. DOI: 10.12697/jeful.2015.6.2.10.

— (2019): *Funktionen von Possessivsuffixen in den ugrischen Sprachen*. Köln: MAP. DOI: 10.16994/bal.

Janda, Gwen Eva & Wisiorek, Axel & Eckmann, Stefanie (2017): Reference Tracking Mechanisms and Automatic Annotation Based on Ob-Ugric Information Structure. *Suomalais-Ugrilaisen Seuran Aikakauskirja. Journal de la Société Finno-Ougrienne* 2017.96, 115–126. DOI: 10.33340/susa.70251.

Johnson-Laird, Philip N. (1980): Mental Models in Cognitive Science. *Cognitive Science* 4.1, 71–115. DOI: 10.1207/s15516709cog0401_4.

— (1983): *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive Science Series 6. Cambridge, MA: Harvard University Press.

Kant, Immanuel (1998): *Kritik der reinen Vernunft*. Hrsg. von Jens Timmermann. Erstausgabe 1781/1787. Hamburg: Meiner.

Karlgren, Jussi & Cutting, Douglass (1994): Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: *Proceedings of the 15th Conference on Computational Linguistics*. Bd. 2. COLING '94. Kyoto: Association for Computational Linguistics, 1071–1075. DOI: 10.3115/991250.991324.

Karlsson, Fred (2007): Constraints on Multiple Center-Embedding of Clauses. *Journal of Linguistics* 43.2, 365–392. DOI: 10.1017/S0022226707004616.

Kassambara, Alboukadel & Mundt, Fabian (2017): *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Dokumentation. URL: https://CRAN.R-project.org/package=factoextra [abgerufen am 07. 08. 2020].

Keresztes, László (1998): Mansi. In: *The Uralic Languages*. Hrsg. von Daniel Abondolo. London & New York: Routledge, 387–427.

Kerezsi, Agnezs (1995): Music Instruments in the Ritual Ceremonies of the Ob-Ugrians. In: *Folk Belief Today*. Hrsg. von Mare Kõiva & Kai Vassiljeva. Tartu: Estonian Academy of Sciences, 182–188. URI: https://www.folklore.ee/rl/pubte/ee/usund/fbt/kerezsi.pdf.

Kessler, Brett & Nunberg, Geoffrey & Schütze, Hinrich (1997): Automatic Detection of Text Genre. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. ACL '98/EACL '98. Madrid: Association for Computational Linguistics, 32–38. DOI: 10.3115/976909.979622.

Kintsch, Walter & van Dijk, Teun A. (1978): Toward a Model of Text Comprehension and Production. *Psychological Review* 85.5, 363–394. DOI: 10.1037/0033-295X.85.5.363.

Kolga, Margus u. a. (2020a): The Khants. In: *The Red Book of the Peoples of the Russian Empire*. URL: https://web.archive.org/web/20210607204516/https://www.eki.ee/books/redbook/khants.shtml [abgerufen am 23. 07. 2020].

— (2020b): The Mansis. In: *The Red Book of the Peoples of the Russian Empire*. URL: https://web.archive.org/web/20210607204501/https://www.eki.ee/books/redbook/mansis.shtml [abgerufen am 23. 07. 2020].

Kopp, Birgitta (2020): Schematheorie. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 1551.

Kopp, Birgitta & Caspar, Franz (2020): Schema. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 1550f.

Krummenacher, Joseph (2020): Aufmerksamkeit. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 221f.

Kuhn, Max u. a. (2018): *caret: Classification and Regression Training*. Dokumentation. URL: https://CRAN.R-project.org/package=caret [abgerufen am 06. 08. 2020].

Kuksa, Pavel P. (2014): Efficient Multivariate Sequence Classification. *Computing Research Repository (CoRR)* 1409, 1–9. URI: https://arxiv.org/abs/1409.8211.

Labov, William (1972): The Transformation of Experience in Narrative Syntax. In: *Language in the Inner City. Studies in the Black English Vernacular*. Hrsg. von William Labov. Philadelphia: University of Philadelphia Press, 354–396.

Lambrecht, Knud (1994): *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge Studies in

Linguistics 71. Cambridge: Cambridge University Press. DOI: 10.1017/CBO 9780511620607.

Landis, J. Richard & Koch, Gary G. (1977): The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33.1, 159–174. DOI: 10.2307/2529310.

Langacker, Ronald W. (1986): An Introduction to Cognitive Grammar. *Cognitive Science* 10.1, 1–40. DOI: 10.1207/s15516709cog1001_1.

— (1991): Cognitive Grammar. In: *Linguistic Theory and Grammatical Description: Nine Current Approaches*. Hrsg. von Flip G. Droste & John E. Joseph. Current Issues in Linguistic Theory 75. Amsterdam: Benjamins, 275–306. DOI: 10.1075/cilt.75.10lan.

— (2000): A Dynamic Usage-Based Model. In: *Usage-Based Models of Language*. Hrsg. von Michael Barlow & Suzanne Kemmer. Stanford, CA: CSLI Publications, 1–63.

Lapshinova-Koltunski, Ekaterina & Zampieri, Marcos (2018): Linguistic Features of Genre and Method Variation in Translation: A Computational Perspective. In: *The Grammar of Genres and Styles*. Hrsg. von Dominique Legallois & Thierry Charnois & Meri Larjavaara. Berlin & Boston: De Gruyter, 92–117. DOI: 10.1515/9783110595864-005.

Legallois, Dominique & Charnois, Thierry & Larjavaara, Meri (2018a): Grammar of Genres and Styles: An Overview. In: *The Grammar of Genres and Styles*. Hrsg. von Dominique Legallois & Thierry Charnois & Meri Larjavaara. Berlin & Boston: De Gruyter, 1–13. DOI: 10.1515/9783110595864-001.

— Hrsg. (2018b): *The Grammar of Genres and Styles*. Berlin & Boston: De Gruyter. DOI: 10.1515/9783110595864.

Leslie, Christina & Eskin, Eleazar & Noble, William S. (2002): The Spectrum Kernel: A String Kernel for SVM Protein Classification. In: *Proceedings of the Pacific Symposium on Biocomputing 2002*. Hrsg. von Russ B. Altman u. a. Singapur: World Scientific, 564–575. DOI: 10.1142/9789812799623_0053.

Lévi-Strauss, Claude (1972): Die Struktur der Mythen. In: *Strukturalismus in der Literaturwissenschaft*. Hrsg. von Heinz Blumensath. Köln: Kiepenheuer & Witsch, 25–46.

Li, Charles N. & Thompson, Sandra A. (1976): Subject and Topic: A New Typology of Language. In: *Subject and Topic*. Hrsg. von Charles N. Li. New York: Academic Press, 457–489.

— (1979): Third-Person Pronouns and Zero-Anaphora in Chinese Discourse. In: *Discourse and Syntax*. Hrsg. von Talmy Givón. Syntax and Semantics 12. New York [u. a.]: Brill, 311–335. DOI: 10.1163/9789004368897_014.

Liaw, Andy & Wiener, Matthew (2018): *Package 'randomForest' – Breiman and Cutler's Random Forests for Classification and Regression*. URL: https://cran.r-

project.org/web/packages/randomForest/randomForest.pdf [abgerufen am 19.05.2020].

Lintrop, Aado (2020): *The Mansi – History and Present Day*. URL: https://web.archive.org/web/20210614140141/http://folklore.ee/~aado/rahvad/mansingl.htm [abgerufen am 08.07.2020].

Liu, Huan & Motoda, Hiroshi (1998): Less Is More. In: *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Hrsg. von Huan Liu & Hiroshi Motoda. Boston, MA: Springer, 3–12. DOI: 10.1007/978-1-4615-5725-8_1.

Longacre, Robert E. (1983): *The Grammar of Discourse*. New York [u. a.]: Plenum Press. DOI: 10.1007/978-1-4615-8018-8.

Lüthi, Max (1998): *Es war einmal: Vom Wesen des Volksmärchens*. Göttingen: Vandenhoeck & Ruprecht.

Mabroukeh, Nizar R. & Ezeife, Christie I. (2010): A Taxonomy of Sequential Pattern Mining Algorithms. *ACM Computing Surveys* 43.1, 1–41. DOI: 10.1145/1824795.1824798.

Mandl, Heinz & Friedrich, Helmut F. & Hron, Aemilian (1988): Theoretische Ansätze zum Wissenserwerb. In: *Wissenspsychologie*. Hrsg. von Heinz Mandl & Hans Spada. München [u. a.]: Psychologie Verlags Union, 123–160.

Manning, Christopher D. & Raghavan, Prabhakar & Schütze, Hinrich (2009): *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. URI: https://nlp.stanford.edu/IR-book.

Manning, Christopher D. & Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press. URI: https://nlp.stanford.edu/fsnlp.

Martin, J. R. (2001): Cohesion and Texture. In: *The Handbook of Discourse Analysis*. Hrsg. von Deborah Schiffrin & Deborah Tannen & Heidi E. Hamilton. Malden, MA: Blackwell, 35–53. DOI: 10.1002/9780470753460.ch3.

Masseglia, Florent & Teisseire, Maguelonne & Poncelet, Pascal (2004): Extraction de motifs séquentiels. Problèmes et méthodes. *Revue des Sciences et Technologies de l'Information – Série ISI: Ingénierie des Systèmes d'Information* 9.3/4, 183–210. URI: https://hal-lirmm.ccsd.cnrs.fr/lirmm-00108563.

Matić, Dejan (2015): Information Structure in Linguistics. In: *The International Encyclopedia of the Social & Behavioral Sciences*. Hrsg. von James D. Wright. Amsterdam: Elsevier, 95–99. DOI: 10.1016/B978-0-08-097086-8.53013-X.

May, Mark (2020): Kognitive Karte. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 949.

Meder, Theo u. a. (2016): Automatic Enrichment and Classification of Folktales in the Dutch Folktale Database. *The Journal of American Folklore* 129, 78–96. DOI: 10.5406/jamerfolk.129.511.0078.

Mehler, Alexander (2005): Eigenschaften der textuellen Einheiten und Systeme. In: *Quantitative Linguistik / Quantitative Linguistics*. Hrsg. von Reinhard Köhler & Gabriel Altmann & Rajmund G. Piotrowski. Handbücher zur Sprach- und Kommunikationswissenschaft 27. Berlin & New York: De Gruyter, 325–348. DOI: 10.1515/9783110155785.

Miller, Carolyn R. (1984): Genre as Social Action. *Quarterly Journal of Speech* 70.2, 151–167. DOI: 10.1080/00335638409383686.

Minsky, Marvin (1974): A Framework for Representing Knowledge. *MIT Artificial Intelligence Memo* 306, 1–81. URI: http://hdl.handle.net/1721.1/6089.

Mitchell, Tom M. (2017): *Key Ideas in Machine Learning*. URL: http://www.cs.cmu.edu/~tom/mlbook/keyIdeas.pdf [abgerufen am 16. 09. 2018].

Mooney, Carl H. & Roddick, John F. (2013): Sequential Pattern Mining – Approaches and Algorithms. *ACM Computing Surveys* 45.2, 1–39. DOI: 10.1145/2431211.2431218.

Mooney, Raymond J. (2004): Machine Learning. In: *The Oxford Handbook of Computational Linguistics*. Hrsg. von Ruslan Mitkov. Oxford: Oxford University Press, 376–394. DOI: 10.1093/oxfordhb/9780199276349.013.0020.

Motoda, Hiroshi & Liu, Huan (2002): Feature Selection, Extraction and Construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan)* 5, 67–72.

Murtagh, Fionn & Legendre, Pierre (2014): Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* 31.3, 274–295. DOI: 10.1007/s00357-014-9161-z.

Myhill, John (2001): Typology and Discourse Analysis. In: *The Handbook of Discourse Analysis*. Hrsg. von Deborah Schiffrin & Deborah Tannen & Heidi E. Hamilton. Malden, MA: Blackwell, 161–174. DOI: 10.1002/9780470753460.ch9.

Needleman, Saul B. & Wunsch, Christian D. (1970): A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* 48.3, 443–453. DOI: 10.1016/0022-2836(70)90057-4.

Nguyen, Dong u. a. (2012): Automatic Classification of Folk Narrative Genres. In: *Proceedings of KONVENS 2012*. LThist 2012 Workshop. Hrsg. von Jeremy Jancsary. Wien: ÖGAI, 378–382. URI: https://konvens.org/proceedings/2012/pdf/56_nguyen12w.

Nikolaeva, Irina (1999): *Ostyak*. Languages of the World/Materials 305. München: Lincom Europa.

— (2001): Secondary Topic as a Relation in Information Structure. *Linguistics* 39.1, 1–49. DOI: 10.1515/ling.2001.006.

Ochsner, Kevin N. (2007): Social Cognitive Neuroscience: Historical Development, Core Principles, and Future Promise. In: *Social Psychology: Handbook of Basic*

Principles. Hrsg. von Arie W. Kruglanski & E. Tory Higgins. New York: Guilford Publications, 39–66. URI: https://psycnet.apa.org/record/2007-11239-003.

Ojamaa, Triinu & Ross, Jaan (2004): Relationship between Texts and Tunes in the Siberian Folksongs. In: *Conference on Interdisciplinary Musicology*. CIM04. Hrsg. von Richard Parncutt & Annekatrin Kessler & Fränk Zimmer. Graz: Universität Graz, 134–135.

OUDB (2017a): *IS-Annotation: Anleitung zur Annotation*. URL: http://www.babel.gwi.uni-muenchen.de/media/downloads/Anleitung_Annotation_05-06-17.pdf [abgerufen am 14. 07. 2021].

— (2017b): *IS-Annotation: Tags und Annotationsregeln*. URL: http://www.babel.gwi.uni-muenchen.de/media/downloads/IS_Tags_und_Annotationsregeln_13-06-17.pdf [abgerufen am 14. 07. 2021].

— (2021): *Semi-Automatic Annotation of Functional, Semantic and Pragmatic Roles for Ob-Ugric Texts*. URL: http://www.babel.gwi.uni-muenchen.de?abfrage=tagging [abgerufen am 14. 07. 2021].

Panzer, Friedrich (2020): *Märchen*. edition amalia. (Elektronische Publikation. Ursprünglich erschienen in: Deutsche Volkskunde. 1926. Hrsg. von John Meier. Berlin & Leipzig: De Gruyter). URL: http://www.maerchenlexikon.de/texte/archiv/panzer01.htm [abgerufen am 08. 07. 2020].

Peng, Yan (2020): *Text, Grammar, and Worlds: Towards a Narrative Typology of Quechua Folk Tales*. Diss. München: Ludwig-Maximilians-Universität München. DOI: 10.5282/edoc.26158.

Petitjean, François & Ketterlin, Alain & Gançarski, Pierre (2011): A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. *Pattern Recognition* 44.3, 678–693. DOI: 10.1016/j.patcog.2010.09.013.

Piaget, Jean (1948): *Psychologie der Intelligenz*. Zürich: Rascher.

Pinker, Steven (1979): Formal Models of Language Learning. *Cognition* 7.3, 217–283. DOI: 10.1016/0010-0277(79)90001-5.

Propp, Vladimir J. (1972): *Morphologie des Märchens*. München: Hanser.

R Core Team (2018): *The R Project for Statistical Computing*. Dokumentation. URL: https://www.R-project.org [abgerufen am 06. 08. 2020].

— (2020): *hclust: Hierarchical Clustering*. Dokumentation. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html [abgerufen am 05. 08. 2020].

Reiter, Nils & Frank, Anette & Hellwig, Oliver (2014): An NLP-Based Cross-Document Approach to Narrative Structure Discovery. *Literary and Linguistic Computing* 29.4, 583–605. DOI: 10.1093/llc/fqu055.

Renner, Karl N. (2000): Die strukturalistische Erzähltextanalyse. In: *Text- und Gesprächslinguistik*. Hrsg. von Klaus Brinker. Bd. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 16. Berlin & New York: De Gruyter, 43–53. DOI: 10.1515/9783110194067.

Riese, Timothy (2001): *Vogul*. Languages of the World/Materials 158. München: Lincom Europa.

Ritschard, Gilbert & Bürgin, Reto & Studer, Matthias (2013): Exploratory Mining of Life Event Histories. In: *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*. Hrsg. von John J. McArdle & Gilbert Ritschard. New York: Routledge, 221–253. DOI: 10.4324/9780203403020-18.

Rousseeuw, Peter J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. DOI: 10.1016/0377-0427(87)90125-7.

Rubin, David C. (1995): *Memory in Oral Traditions. The Cognitive Psychology of Epic, Ballads, and Counting-Out Rhymes*. New York & Oxford: Oxford University Press. URI: https://psycnet.apa.org/record/1995-97902-000.

Rumelhart, David E. (1975): Notes on a Schema for Stories. In: *Representation and Understanding: Studies in Cognitive Science*. Hrsg. von Daniel G. Bobrow & Allan Collins. New York: Academic Press, 211–236. DOI: 10.1016/B978-0-12-108550-6.50013-6.

Rumelhart, David E. & Hinton, Geoffrey E. & Williams, Ronald J. (1986): Learning Representations by Back-Propagating Errors. *Nature* 323, 533–536. DOI: 10.1038/323533a0.

Rumelhart, David E. & Ortony, Andrew (1977): The Representation of Knowledge in Memory. In: *Schooling and the Acquisition of Knowledge*. Hrsg. von Richard C. Anderson & Rand J. Spiro & William E. Montague. London & New York: Routledge, 99–135. DOI: 10.4324/9781315271644-10.

Ryan, Marie-Laure (2003): Cognitive Maps and the Construction of Narrative Space. In: *Narrative Theory and the Cognitive Sciences*. Hrsg. von David Herman. CSLI Lecture Notes 158. Stanford, CA: CSLI Publications, 214–242.

Sakoe, Hiroaki & Chiba, Seibi (1978): Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 26.1, 43–49. DOI: 10.1109/TASSP.1978.1163055.

Sanders, Ted & Spooren, Wilbert (2007): Discourse and Text Structure. In: *The Oxford Handbook of Cognitive Linguistics*. Hrsg. von Dirk Geeraerts & Hubert Cuyckens. Oxford: Oxford University Press, 916–941. DOI: 10.1093/oxfordhb/9780199738632.013.0035.

Sarda-Espinosa, Alexis (2018): *dtwclust: Time Series Clustering along with Optimizations for the Dynamic Time Warping Distance*. Dokumentation. URL: https://CRAN.R-project.org/package=dtwclust [abgerufen am 07. 08. 2020].

Sauer, Gert (2004–2005): Formeln und Formelhaftes in ostjakischen Märchen. *Finnisch-Ugrische Mitteilungen* 28-29, 331–338.

Schiffrin, Deborah (1981): Tense Variation in Narrative. *Language* 57.1, 45–62. DOI: 10.2307/414286.

Schnedecker, Catherine (2018): Reference Chains and Genre Identification. In: *The Grammar of Genres and Styles*. Hrsg. von Dominique Legallois & Thierry Charnois & Meri Larjavaara. Berlin & Boston: De Gruyter, 39–66. DOI: 10.1515/9783110595864-003.

Schön, Zsófia (2017): *Postpositionale Konstruktionen in chantischen Dialekten*. Dissertationen der LMU 15. München: Universitätsbibliothek der Ludwig-Maximilians-Universität München. DOI: 10.5282/edoc.20716.

Schulze, Wolfgang (2000a): The Cognitive Dimension of Clausal Organization in Udi. In: *First International Conference on Cognitive Typology*. Antwerpen. URI: http://www.schulzewolfgang.de/material/Udicog.pdf [abgerufen am 26. 04. 2020].

— (2000b): Towards a Typology of the Accusative Ergative Continuum: The Case of East Caucasian. *General Linguistics* 37.1, 71–155.

— (2004a): Pragmasyntax: Towards a Cognitive Typology of the Attention Information Flow in Udi Narratives. In: *Linguagem, Cultura e Cognição: Estudos de Linguística Cognitiva*. Hrsg. von Augusto Soares da Silva & Amadeu Torres & Miguel Gonçalves. Coimbra: Almedina, 545–574.

— (2004b): Rezension: Nikolaeva, Irina & Perekhvalskaya, Elena & Tolskaya, Maria. Udeghe (Udihe) Folk Tales. 2002. *Studies in Language* 28.1, 203–208. DOI: 10.1075/sl.28.1.09sch.

— (2011): *Cognitive Transitivity. The Motivation of Basic Clause Structures*. URL: https://www.academia.edu/1802287/Cognitive_Transivity [abgerufen am 30. 12. 2014].

— (2013): *Eine kognitive Typologie der Zeit. Prolegomenon*. URL: http://www.schulzewolfgang.de/material/kgzeit.pdf [abgerufen am 26. 04. 2020].

— (2018): *Schemas, Models, and Constructions. The Linguistic Representation of Event Image Structures in Grammar and Narration*. URL: https://web.archive.org/web/20210713110110/http://www.schulzewolfgang.de/temp/schemas_models_draft_1.pdf [abgerufen am 13. 07. 2021].

— (2019): *Genre-gesteuerte linguistische Praktiken. Text-Weltmodelle von Volkserzählungen*. (Manuskript; in modifizierter Fassung erschienen 2020).

— (2020): Explorationen zur Genre-Grammatik von Volksnarrationen. *Zeitschrift für Germanistische Linguistik* 48.3, 590–636. DOI: 10.1515/zgl-2020-2015.

Schulze, Wolfgang & Sallaberger, Walther (2007): Grammatische Relationen im Sumerischen. *Zeitschrift für Assyriologie und Vorderasiatische Archäologie* 97.2, 163–214. DOI: 10.1515/ZA.2007.010.

Schütz, Alfred & Luckmann, Thomas (2003): *Strukturen der Lebenswelt*. Konstanz: UVK.

Schütze, Hinrich (1997): *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. CSLI Lecture Notes 71. Stanford, CA: CSLI Publications.

Schütze, Hinrich & Hull, David A. & Pedersen, Jan O. (1995): A Comparison of Classifiers and Document Representations for the Routing Problem. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '95. Seattle, WA: Association for Computing Machinery, 229–237. DOI: 10.1145/215206.215365.

Schwarz, Monika (2000): *Indirekte Anaphern in Texten: Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Linguistische Arbeiten 413. Tübingen: Niemeyer. DOI: 10.1515/9783110912517.

— (2001): Establishing Coherence in Text. Conceptual Continuity and Text-world Models. *Logos and Language* 2.1, 15–24.

Schwarz-Friesel, Monika & Consten, Manfred (2011): Reference and Anaphora. In: *Foundations of Pragmatics*. Hrsg. von Wolfram Bublitz & Neal R. Norrik. Berlin & New York: De Gruyter, 347–372. DOI: 10.1515/9783110214260.347.

— (2014): *Einführung in die Textlinguistik*. Darmstadt: WBG.

Seitz, Daniel (2020): Arbeitsgedächtnis. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 186.

Siikala, Anna-Leena & Ulyashev, Oleg (2011): *Hidden Rituals and Public Performances. Traditions and Belonging among the post-Soviet Khanty, Komi and Udmurts*. Studia Fennica Folkloristica 19. Helsinki: Finnish Literature Society. DOI: 10.21435/sff.19.

Sim, Julius & Wright, Chris C. (2005): The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* 85.3, 257–268. DOI: 10.1093/ptj/85.3.257.

Skopeteas, Stavros u. a. (2006): *Questionnaire on Information Structure: Reference Manual*. Bd. 4. Interdisciplinary Studies on Information Structure. Potsdam: Universitätsverlag Potsdam. URI: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/information-structure_description.php.

Skribnik, Elena (2001): Pragmatic Structuring in Northern Mansi. In: *Congressus Nonus Internationalis Fenno-Ugristarum. 7.–13.8.2000 Tartu. Pars VI. Dissertationes sectionum: Linguistica III*. Hrsg. von Tõnu Seilenthal. Tartu: Auctores, 222–239.

— (2010): Hierarchy Effects in Northern Mansi. In: *RHIM Meeting*. Leipzig. URI: http://www.babel.gwi.uni-muenchen.de/media/downloads/grammar/NorthernMansi/Syntax/hierarchy_effects_northern_mansi.pdf [abgerufen am 04.01.2017].

Starfield, Tony (2005): *Principles of Modeling: Real World – Model World*. URL: http://www.uvm.edu/~tdonovan/modeling/Module1/01_RealWorld-ModelWorld_transcript.pdf [abgerufen am 21.04.2020].

Studer, Matthias (2013): WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R. *LIVES Working Papers* 24, 1–32. DOI: 10.12682/lives.2296-1658.2013.24.

Studer, Matthias & Ritschard, Gilbert (2016): What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.2, 481–511. DOI: 10.1111/rssa.12125.

Stukker, Ninke & Spooren, Wilbert & Steen, Gerard (2016): Genre in Language, Discourse and Cognition: Introduction to the Volume. In: *Genre in Language, Discourse and Cognition*. Hrsg. von Ninke Stukker & Wilbert Spooren & Gerard Steen. Applications of Cognitive Linguistics 33. Berlin & Boston: De Gruyter, 1–12. DOI: 10.1515/9783110469639-002.

Swales, John M. (1990): *Genre Analysis: English in Academic and Research Settings*. Cambridge [u. a.]: Cambridge University Press.

Tan, Pang-Ning & Steinbach, Michael & Kumar, Vipin (2006): *Introduction to Data Mining*. Boston: Pearson.

Thomaschke, Roland (2020): Zeit. In: *Dorsch – Lexikon der Psychologie*. Hrsg. von Markus A. Wirtz & Friedrich Dorsch. 19. Aufl. Bern: Hogrefe, 1954.

Thompson, Chad L. (1989): Voice and Obviation in Navajo. In: *Proceedings of the Fourth Annual Meeting of the Pacific Linguistics Conference*. Hrsg. von Robert Carlson u. a. Eugene, OR: University of Oregon, 466–488.

Thompson, James R. (2011): *Empirical Model Building: Data, Models, and Reality*. Hoboken: Wiley. DOI: 10.1002/9781118109656.

Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Studies in Corpus Linguistics 6. Amsterdam & Philadelphia: Benjamins. DOI: 10.1075/scl.6.

Toussaint, Marc (2003): Learning a World Model and Planning with a Self-Organizing, Dynamic Neural System. In: *Advances in Neural Information Processing Systems 16*. NIPS 2003. Hrsg. von Sebastian Thrun & Lawrence K. Saul & Bernhard Schölkopf. Cambridge, MA: MIT Press, 929–937. URI: https : / / proceedings . neurips . cc / paper / 2003 / hash / 28b60a16b55fd531047c0c958ce14b95-Abstract.html.

Tukey, John W. (1977): *Exploratory Data Analysis*. Reading, MA [u. a.]: Addison-Wesley.

UNESCO (2021): *Arts of the Meddah, Public Storytellers*. Inscribed in 2008 (3.COM) on the Representative List of the Intangible Cultural Heritage of Humanity (originally proclaimed in 2003). Türkiye. URL: https://ich.unesco.org/en/RL/arts-of-the-meddah-public-storytellers-00037 [abgerufen am 21. 10. 2021].

Ungerer, Friedrich & Schmid, Hans-Jörg (1996): *An Introduction to Cognitive Linguistics*. London [u. a.]: Longman.

Uther, Hans-Jörg (2011): *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson. Part 1: Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction*. Bd. 1. FF Communications 284. Helsinki: Academia Scientiarum Fennica.

Van der Auwera, Johan & Nuyts, Jan (2007): Cognitive Linguistics and Linguistic Typology. In: *The Oxford Handbook of Cognitive Linguistics*. Hrsg. von Dirk Geeraerts & Hubert Cuyckens. Oxford: Oxford University Press, 1074–1091. DOI: 10.1093/oxfordhb/9780199738632.013.0040.

Van Valin, Robert D. & LaPolla, Randy J. (1997): *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139166799.

Van Dijk, Teun A. (1972): *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics*. Janua Linguarum. Series Maior 63. Den Haag & Paris: De Gruyter. DOI: 10.1515/9783110804263.

— (1978): Aspekte einer Textgrammatik. In: *Textlinguistik*. Hrsg. von Wolfgang U. Dressler. Darmstadt: WBG, 268–299.

— (2007): Editor's Introduction: The Study of Discourse: An Introduction. In: *Discourse Studies*. Hrsg. von Teun A. van Dijk. Bd. 1. SAGE Benchmarks in Discourse Studies. London: SAGE, xix–xlii.

— (2016): Critical Discourse Studies: A Sociocognitive Approach. In: *Methods of Critical Discourse Analysis*. Hrsg. von Ruth Wodak & Michael Meyer. London: SAGE, 63–85.

— (2018): Sociocognitive Discourse Studies. In: *The Routledge Handbook of Critical Discourse Studies*. Hrsg. von John Flowerdew & John E. Richardson. London & New York: Routledge, 26–43. DOI: 10.4324/9781315739342-3.

Van Dijk, Teun A. & Kintsch, Walter (1978): Cognitive Psychology and Discourse: Recalling and Summarizing Stories. In: *Current Trends in Textlinguistics*. Hrsg. von Wolfgang U. Dressler. Research in Text Theory 2. Berlin: De Gruyter, 61–80. DOI: 10.1515/9783110853759.61.

— (1983): *Strategies of Discourse Comprehension*. New York [u. a.]: Academic Press.

Virtanen, Susanna (2014): Pragmatic Direct Object Marking in Eastern Mansi. *Linguistics* 52.2, 391–413. DOI: 10.1515/ling-2013-0067.

— (2015): *Transitivity in Eastern Mansi: An Information Structural Approach*. Diss. Helsinki: University of Helsinki. URI: http://urn.fi/URN:ISBN:978-951-51-0548-6.

Virtanen, Susanna & Sosa, Sachiko (2018): Remarks on Areal Linguistics in the Information Structure of the Ob-Ugrian Languages. *Folia Uralica Debreceniensia* 25, 233–260. URI: https://researchportal.helsinki.fi/en/publications/afbc0009-676c-4d91-af12-d074d9ba496d.

Virtanen, Tuija (2009): Corpora and Discourse Analysis. In: *Corpus Linguistics*. Hrsg. von Anke Lüdeling & Merja Kytö. Bd. 2. Handbücher zur Sprach- und Kommunikationswissenschaft 29. Berlin & New York: De Gruyter, 1043–1070. DOI: 10.1515/9783110213881.2.1043.

Ward, Joe H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58.301, 236–244. DOI: 10.1080/01621459.1963.10500845.

Weinrich, Harald (1976): *Sprache in Texten*. Stuttgart: Klett.

Werlich, Egon (1975): *Typologie der Texte: Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Heidelberg: Quelle & Meyer.

Werth, Paul (1999): *Text Worlds: Representing Conceptual Space in Discourse*. London [u. a.]: Longman.

Wisiorek, Axel & Schön, Zsófia (2017): Ob-Ugric Database: Corpus and Lexicon Databases of Khanty and Mansi Dialects. *Acta Linguistica Academica* 64.3, 383–396. DOI: 10.1556/2062.2017.64.3.4.

Xing, Zhengzheng & Pei, Jian & Keogh, Eamonn (2010): A Brief Survey on Sequence Classification. *ACM SIGKDD Explorations Newsletter* 12.1, 40–48. DOI: 10.1145/1882471.1882478.

Zehetmaier, Marianne & Fónyad, Gábor (2020): Waldgeist. In: *Ethnographical Comments (Ob-Ugric Languages: Conceptual Structures, Lexicon, Constructions, Categories)*, 45f. URL: http://www.babel.gwi.uni-muenchen.de/media/downloads/ethno_comm_dt.pdf [abgerufen am 23. 07. 2020].

Zwaan, Rolf A. & Radvansky, Gabriel A. (1998): Situation Models in Language Comprehension and Memory. *Psychological Bulletin* 123.2, 162–185. DOI: 10.1037/0033-2909.123.2.162.

Vorliegende Arbeit verbindet automatische Verfahren der Mustererkennung und der explorativen Feature-Analyse mit textlinguistischen Parametern einer kognitiven Texttypologie, um eine Methodik für eine kognitiv adäquate, gebrauchsbasierte Genre-Klassifizierung anhand von annotierten Korpusdaten zu entwickeln. Zu den hier relevanten Parametern zählen, neben einfachen textstatistischen Maßen mit kognitiver Interpretation als Elaborationsmaße, vor allem Merkmale des referentiellen, relationalen sowie informationsstrukturellen Aufbaus textuell kodierter kognitiver Modelle, wie referentielle Distanz, häufige Ereignisschemata, Informationsdichte oder Muster textinterner Diskursstrukturierung. Durch Anwendung von Klassifikations- und Clusteringalgorithmen auf ein zeitlich und dialektal geschichtetes, syntaktisch, semantisch und informationsstrukturell annotiertes Korpus obugrischer Volkserzählungen sowie weiterer, primär mündlicher Genres wird die Eignung dieser Methodik einer automatischen Induktion quantitativer Textstrukturtypen für die Rekonstruktion von Text-Weltmodellen als genrespezifischen, durch Typisierung von Sprachgebrauchssituationen erlernten, schematischen Textstruktur-Modellen der menschlichen Kognition evaluiert.

Axel Wisiorek studierte Philosophie, Allgemeine Sprachwissenschaft sowie Computerlinguistik an der Ludwig-Maximilians-Universität München, wo er seit 2014 auch als wissenschaftlicher Mitarbeiter tätig ist. 2021 promovierte er dort in Allgemeiner Sprachwissenschaft mit Forschungsschwerpunkt auf der Verbindung kognitiv begründeter Sprachtypologie mit Methoden des maschinellen Lernens in der Korpuslinguistik.

84,00 €
ISBN 978-3-487-16116-7