LEMoN

Léxico de Expresiones Multipalabra Nominales

Ein spanisches Lexikon nominaler Mehrworteinheiten,
entwickelt mit lokalen Grammatiken

Inaugural-Dissertation

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

vorgelegt von
Daniel Stein aus Kassel

2026

Erstgutachter: **PD Dr. Stefan Langer**
(Ludwig-Maximilians-Universität München)

Zweitgutachter: **Prof. Dr. phil. Andreas Dufter**
(Ludwig-Maximilians-Universität München)

Tag der mündlichen Prüfung: 06.02.2026

Tú que me lees, ¿estás seguro de entender mi lenguaje?
- Jorge Luis Borges

Meiner Mutter gewidmet.

Danksagung

Auch wenn es sich manchmal so anfühlt: Als Promovend und als Lexikograph ist man selten allein mit den Wörtern. Ich möchte die Gelegenheit nutzen, denen zu danken, die mir auf dem Weg geholfen haben.

Mein Dank gilt insbesondere meinem Doktorvater, PD Dr. Stefan Langer, für die Übernahme meiner Arbeit in einer schwierigen Phase - und für die ausdauernde Betreuung. Auch meinem Zweitbetreuer, Prof. Dr. Andreas Dufter, danke ich für die Unterstützung und dafür, dass er die spanische Perspektive in der Arbeit gestärkt hat. *¡Muchas gracias!*

Ebenso möchte ich mich bei Prof. Dr. Franz Guenthner und Prof. Dr. Xavier Blanco Escoda bedanken, die mich zu einer Promotion im Umfeld der Gross-Schule inspiriert und in der ersten Phase engagiert begleitet haben. Besonders dankbar bin ich für die Möglichkeit eines Forschungsaufenthalts in Barcelona, von dem Arbeit und Autor spürbar profitiert haben.

Weiterer Dank gilt meinen Kolleginnen und Kollegen sowie meinen befreundeten Wissenschaftlerinnen und Wissenschaftlern, für offene Ohren und kluge Ratschläge.

Meinen Freundinnen und Freunden herzlichen Dank für die moralische Unterstützung über viele Jahre - und zuletzt beim Korrekturlesen.

Schließlich danke ich meinem Vater, meinem Bruder und meiner ganzen Familie für Geduld, Verständnis und Rückhalt über die gesamte Entstehungszeit.

Im Gedenken an die Angehörigen, denen ich nicht mehr persönlich danken kann.

Mein besonderer Dank gilt meiner Frau und unseren Töchtern. Ohne Euch hätte ich das nicht geschafft.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ASALE Asociación de Academias de la Lengua Española 69

CIS Centrum für Informations- und Sprachverarbeitung 59

CN Compound Noun 49

CL Computerlinguistik 21

DELA Dictionnaire Electronique du LADL 98

DiCE Diccionario de Colocaciones Españoles 60

FSA Finite State Automaton 91

FST Finite State Transducer 91

GUI Graphical User Interface 114

IR Information Retrieval 21

IE Information Extraction 21

ISO International Standards Organization 111

KI Künstliche Intelligenz 20

KWIC Keyword in Context 90

LADL Laboratoire d'Automatique Documentaire et Linguistique 58

LEMoN Lexicón Español de Expresiones Multipalabras Nominales 22

LLM Large Language Model 56

LMF Lexicon Markup Framework 110

LMU Ludwig-Maximilians-Universität München 59

LVC Light Verb Construction 52

MI Main Item 112

MRD Machine Readable Dictionary 98

MTD Machine Traceable Dictionary 98

MT Machine Translation 21

MWE Mehrworteinheit 20

NGLE Nueva gramática de la lengua española 70

NLP Natural Language Processing 21

NP Nominalphrase 45

OOV Out of Vocabulary 119

PP Präpositionalphrase 163

POS Part of Speech 74

QA Question Answering 21

RAE Real Academia Española 69

RTN Recursive Transition Network 91

SVC Support Verb Construction 52

UAB Universitat Autònoma de Barcelona 60

XML eXtended Markup Language 110

Teil I

Grundlagen

1 Einleitung

Like all other scientists, linguists wish they were physicists.

Joseph D. Becker

1.1 Problemstellung

Während die Naturwissenschaften längst systematisch katalogisieren - ob Teilchen, Käfer oder Gene -, fehlt in Teilen der Linguistik ein vergleichbar umfassendes Vorgehen. Uns fehlt immer noch ein Inventar der lexikalischen Grundeinheiten der Sprache - vielleicht sogar Konsens darüber, was diese Grundeinheiten sind.

Vorläufig sei festgehalten:

Es gibt „einfache Wörter“, die einfache lexikalische Einheiten darstellen. Und es gibt „Gruppen von einfachen Wörtern“, die sich anders verhalten, als reguläre „Gruppen von einfachen Wörtern“. Daher müssen sie (nicht nur) in der Sprachverarbeitung anders behandelt werden. Zum Beispiel müssen sie häufig - trotz ihrer *Mehrwortigkeit* - als syntaktische *Einheit* behandelt werden. Wie sich jedoch reguläre Gruppen von einfachen Wörtern von solchen *Mehrwort-Einheiten* (MWE) vollständig automatisch unterscheiden lassen, ist bis heute ungelöst - und hoch relevant:

> Despite recent advances, including the advent of large (and small) language models, MWEs' inherent complexity and distributional properties continue to impede their effective processing.
>
> [V. Mititelu 2025, 1]

Mit anderen Worten gesagt:

> Identifying the lexical units of text, both simple and compound, is the first step towards its automatic processing.
>
> [J. Baptista 2003, 1]

„Large (and small) language models“ referenziert unter anderem auf moderne Chatsysteme wie *ChatGPT 5*[1]. Diese sind zwar beeindruckende Beispiele für

[1] Ein KI-gestützter Chatbot von OpenAI, Release 2025. Genannt als Beispiel, vergleichbare Systeme zeigen ähnliche Grenzen.

Künstliche Intelligenz (KI), doch auch ihnen mangelt es noch an linguistischem Verständnis, insbesondere in den Feinheiten. Nicht-kompositionale, konventionalisierte Einheiten - also Mehrworteinheiten (MWEs) - bereiten weiterhin erhebliche Schwierigkeiten, vgl. [W. He 2025], [B. Matheny 2025], [D. Phelps 2024], [J. Zhou 2024], etc.

Ein Grund dafür ist die Dominanz des intuitiven Wortbegriffs, die das Problem der MWEs verdeckt.

Ein hilfreicher Vergleich ist die (lange unterschätzte) Größenordnung lexikalischer Ambiguität: Sprachverarbeitung ist seit Beginn des Computerzeitalters eine Kerndisziplin, und rückblickend wirkt die frühe Hoffnung, maschinelle Übersetzung sei in wenigen Jahren gelöst, naiv. Ein wesentlicher Grund für diese Fehleinschätzung war die unterschätzte Mehrdeutigkeit in Texten: Menschen nehmen abwegige Lesarten selten wahr, aber Parser legen sie offen. So ermittelte ein System für den Satz „Hinter dem Betrug werden die gleichen Täter vermutet, die während der vergangenen Tage in Griechenland gefälschte Banknoten in Umlauf brachten.“ 92 Lesarten (vgl. [K.-U. Carstensen 2009, 308]). Diese unerwartet hohe Zahl zeigt, dass etwas, nur weil es uns Menschen selbstverständlich scheint, sich teilweise erst unter maschineller Analyse als erklärungsbedürftig erweist. Genauso verhält es sich bei MWEs: Das starke, intuitive Wortkonzept lässt sie wie einfache Wörter erscheinen - eine Art *Wortmimikry*.

Zwei Wege bieten sich an, dieser Mimikry zu begegnen:

(a) datengetriebene/statistische Verfahren zur automatischen Identifikation von MWEs,

(b) ressourcenbasierte Verfahren, die das Inventar einer Sprache systematisch erfassen und kuratieren.

Der zweite Weg, also die Konstruktion umfassender Datensätze, ist aufwändig, zeit- und arbeitsintensiv und linguistisch anspruchsvoll. Entsprechend sind MWEs in vielen Ressourcen nur unzureichend erfasst bzw. fehlen ganz. Maurel und Guenthner haben 2005 pointiert Folgendes formuliert:

> Among the many astonishing lacunae in current natural language processing research (and even more so in current applications as well) is the more or less complete absence of the notion of a dictionary of polylexical units, in particular of dictionaries of polylexical nouns; [...]
>
> [D. Maurel 2005, 21 f]

Zwei Jahrzehnte später hat sich die Lage verbessert, aber nicht grundlegend verändert: Eine aktuelle Studie über lexikalische Ressourcen, die MWEs abbilden oder einschließen, untersucht 66 Wörterbuchprojekte, darunter 6 spanische (teils mit Fokus auf Lateinamerika), vgl. [V. Mititelu 2025]. Der Fokus der untersuchten Ressourcen liegt klar auf verbalen MWEs - nominale sind deutlich unterrepräsentiert (ca. 7 %). Das offenbart einen nach wie vor fragmentierten Status.

Diese Arbeit folgt dem zweiten Weg und adressiert eine zentrale Lücke: Für das Spanische existiert bislang soweit ersichtlich[2] kein maschinenlesbares, umfangreiches Lexikon nominaler MWEs[3]. Das ist kein Randbefund: Ohne ein verlässliches Inventar dieser Bausteine der Sprache stehen alle Natural Language Processing (NLP)-Tasks, die die Erkennung, Verarbeitung oder Generierung von MWEs beinhalten, vor Schwierigkeiten. Vor diesem Hintergrund ist die Erschließung nominaler MWEs keine Fleißaufgabe, sondern eine zentrale Voraussetzung für präzisere, reproduzierbare NLP-Anwendungen (z. B. Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT) und Question Answering (QA)):

> By providing more natural human-machine interfaces, and more sophisticated access to stored information, language processing has come to play a central role in the multilingual information society.
>
> [S. Bird 2009, IX]

Zugleich besteht ein klar artikulierter Community-Bedarf. Das PARSEME-Positionspapier betont die Notwendigkeit von MWE-Wörterbüchern als Basis für Annotation, Extraktion und Evaluierung und fordert deren systematischen Aufbau (vgl. [A. Savary 2019]). Die vorliegende Arbeit setzt hier an.

1.2 Eingliederung in den wissenschaftlichen Diskurs

Diese Dissertation ist in Computerlinguistik (CL) und Computerlexikographie angesiedelt. Sie knüpft an den distributionellen Ansatz von Zellig S. Harris und die darauf aufbauenden lexikon-grammatischen Arbeiten von Maurice Gross an. Im Zentrum steht die Frage, wie sich Bedeutung aus wiederkehrenden sprachlichen Mustern und deren Verteilung in Texten erschließen lässt. Vor diesem Hintergrund wird ein

[2] Recherchestand: Oktober 2025.

[3] Dieses Defizit ist Teil eines umfassenderen Problems der Ressourcenausstattung des Spanischen insgesamt, das in Kapitel 3.1.2 näher beleuchtet wird.

Verfahren zur computergestützten Erschließung spanischer nominaler MWEs entwickelt.

Maurice Gross hat das Phänomen der MWEs früh und prägend in seiner ganzen Breite beschrieben. Mit der „Lexikon-Grammatik“ und dem Formalismus „lokaler Grammatiken“ schuf er Verfahren, die der Analyse und Verarbeitung komplexer sprachlicher Muster gewachsen sind: Auf Basis empirischer Korpusbelege lassen sich präzise Regeln formulieren, die das Vorkommen von MWEs beschreiben. Auf diesen Ansätzen aufbauend präsentiere ich in dieser Dissertation die Pipeline Lexicón Español de Expresiones Multipalabras Nominales (LEMoN) zur Erstellung eines elektronischen Lexikons spanischer nominaler Mehrworteinheiten (MWEs). Im Folgenden werden die Begriffe „Lexikon“ und „Wörterbuch“ synonym verwendet.

Die Arbeit steht damit in der Tradition des linguistischen Distributionalismus nach Harris und Gross. Letzterer formuliert das Fernziel einer Methodologie, die Bedeutung systematisch aus der Verteilung sprachlicher Muster erschließt:

> The aim of corpus processing is the study of the distribution of meaning in texts. The analysis we propose here consists of a preliminary series of steps in this direction. The ultimate goal supposes that the problem of identification and localization of meaning has been solved. Linguistics is far from nearing such a stage. However, a realistic linguistic methodology based on Z. S. Harris' theories of language is available and needs only to be applied in order to reach full coverage for most languages.
>
> [M. Gross 1995, 8]

Vor diesem Hintergrund wird das zentrale Defizit sichtbar, das die vorliegende Arbeit adressiert.

1.3 Beitrag und Aufbau der Arbeit

Im Fokus steht ein recall-orientierter Ansatz zur Erschließung nominaler MWEs im Spanischen, d. h. eine Methodik, die systematisch auf hohe Abdeckung zielt und Präzision nachordnet.[4]Auf dieser Basis leistet die Arbeit fünf Beiträge, die sowohl Ressource als auch Methodik betreffen:

[4]In Anlehnung an die Metriken des Information Retrieval bezeichnet *Recall* das Verhältnis der korrekt erkannten relevanten Einheiten zur Gesamtzahl relevanter Einheiten, während *Precision* das Verhältnis der korrekt erkannten relevanten Einheiten zu allen erkannten Einheiten angibt (vgl. [C. D. Manning 2008, 154f]).

- **LEMoN-Grammatiken:** Lokale Grammatiken zur Extraktion nominaler MWEs aus spanischen Korpora.
- **LEMoN-Pipeline:** Verfahren zur Erstellung eines umfangreichen Kandidateninventars spanischer nominaler MWEs.
- **LEMoN.dic:** Wörterbuch im DELA-Format, das aus dem Kandidateninventar hervorgeht (ohne manuelle Kuratierung).
- **Hauptlauf-Analyse:** Quantitative und qualitative Auswertung von `LEMoN.dic`.
- **Ablationsstudie:** Evaluation zentraler Designentscheidungen für Pipeline und Grammatiken.

Diese Arbeit ist dreigeteilt:

Im *theoretischen Teil* skizziere ich das Phänomen der MWEs mit Terminologie, Klassifikationen und einem Abriss der Forschungsgeschichte. Der Fokus liegt auf der Gross-Schule. Ihre lokalen Grammatiken bilden die Grundlage der hier entwickelten LEMoN-Pipeline sowie der Taxonomien spanischer nominaler MWEs. Außerdem gebe ich einen Überblick über gängige Verfahren zur Erkennung und Extraktion.

Im *methodischen Teil* führe ich die für die Umsetzung relevanten Bausteine zusammen: das formale Instrumentarium mit dem DELA-Wörterbuchformat und der praktischen Spezifikation lokaler Grammatiken, die verwendeten Ressourcen aus Korpora und Wörterbüchern sowie die Werkzeuge für die Implementierung mit Unitex.

Im *praktischen LEMoN-Teil* erläutere ich Aufbau und Implementierung der LEMoN-Pipeline und der zugehörigen Grammatiken und führe die Evaluation des Wörterbuchs, der Pipeline und der zentralen Designentscheidungen durch.

2 Mehrworteinheiten

Por eso cada palabra dice lo que dice y además más y otra cosa.

Alejandra Pizarnik

Im Anschluss an die in der Einleitung umrissene Problemstellung werden im Folgenden die theoretischen Grundlagen und Begriffsbestimmungen zu Mehrworteinheiten erläutert.

2.1 Begriffsbestimmung

Die linguistische Verarbeitung von Textdaten basiert zu wesentlichen Teilen auf Wörterbüchern, die möglichst viele Wörter der betreffenden Sprache(n) enthalten. Wörter werden intuitiv als die atomaren bedeutungstragenden Einheiten begriffen (im Gegensatz zu Morphemen, die in der Regel nicht atomar sind, sondern fast immer Teil von Wörtern). Die Identifikation von Wörtern lässt sich dabei leicht anhand ihrer Abgrenzung voneinander durch Leerzeichen, Interpunktionszeichen und bestimmten Sonderzeichen vornehmen. Diese Form der Segmentierung von Wörtern ist historisch bedingt und bringt einige logische Brüche mit sich, die die automatische Verarbeitung erschweren. Diese logischen Brüche lassen sich für verschiedene Sprachen mühelos beispielhaft darstellen.

Im Deutschen gibt es beispielsweise Partikelverben, die zwar im Infinitiv ohne Leerzeichen geschrieben werden, sich aber in bestimmten Konstruktionen nicht nur aufteilen, sondern über weite Teile des Satzes verteilen können.

1) *Er geht für Oma einkaufen.*

2) *Er kaufte für Oma zwar nicht vor jedem, aber doch vor den meisten Wochenenden, ein bisschen ein.*

Im Französischen zeigen Objektklitika ein kontrastives Verhalten: In finiten Formen stehen sie proklitisch (teils mit Apostroph), im affirmativen Imperativ sind sie enklitisch und werden mit Bindestrichen angeschlossen:

1) *Il me l'a donné.*

2) *Donne-le-moi.*

Anhand dieser Beispiele[5] lässt sich bereits belegen, dass Leerzeichen oder bestimmte Satzzeichen keineswegs immer eindeutige Indikatoren für die Grenzen zwischen Wörtern darstellen - zumindest dann nicht, wenn mit einem Wort (vorläufig definiert) eine zusammengehörende Sinneinheit bezeichnet werden soll. Denn zum einen gibt es offenbar zusammengehörende Wörter, die in bestimmten Fällen durch Leerzeichen (und unter Umständen sogar andere Wörter) getrennt werden. Und zum anderen gibt es Wörter, die üblicherweise alleine vorkommen, in bestimmten Fällen aber ihre trennenden Leerzeichen verlieren. Das ist unter anderem begründet in dem historischen Prozess, den Sprachgebrauch und -analyse im Zuge der Verschriftlichung von Lauten und der Herausbildung normierter Schriftsprachen durchlaufen haben: Eine Segmentierung von geschriebenem Text in einzelne Wörter ist für Latein und Griechisch (und auch für das frühmittelalterliche Germanisch) zunächst unüblich (man schreibt die sogenannte *scriptio continua*) und etwa erst ab dem 7. Jahrhundert regelmäßig zu finden (vgl. [P. Saenger 1997]). Sprachen wie Chinesisch, Japanisch oder Khmer markieren Wortgrenzen nicht durch Leerzeichen, sondern erfordern eine kontextbasierte Segmentierung.

Auch innerhalb der einzelnen Sprachen haben sich zum Teil große Unterschiede bei der Segmentierung ergeben, die man von daher eher als Segmentierungstradition denn als logische und formal einwandfreie Einteilung eines Textes in seine atomaren Einheiten deuten sollte. Und zusätzlich zu den verschiedenen Segmentierungstraditionen, die in den Einzelsprachen Anwendung gefunden haben, sind weitere denkbar. Die Segmentierung von Texten in Wörter ist demzufolge zwar motiviert, aber weder zwingend noch formal schlüssig und in ihrer konkreten Ausgestaltung häufig arbiträr.

Die in den Disziplinen Phraseologie und Parömiologie beschriebenen Phänomene, das heißt stehende Wendungen oder Sprichwörter, sind in die obigen Betrachtungen noch nicht eingeflossen. Gleichwohl stellen sie Probleme für die automatische Sprachverarbeitung dar, denn es handelt sich um Wortgruppen, die gemeinsam etwas anderes bedeuten, als die Summe ihrer Teile vermuten ließe, zum Beispiel das oft zitierte *kick the bucket*. Damit stellen sie die vorläufige Definition von „zusammengehörenden Sinneinheiten“ auf die Probe.

Ebenfalls als logischen Bruch begreifen lassen sich Ausdrücke, die so häufig kombiniert werden, dass sie als Einheit angesehen werden können, obwohl sie oberflächlich betrachtet regelmäßige Bildungen sind. Hier lassen sich verschiedene Fixierungs-

[5] Auch im Spanischen gibt es entsprechende Fälle (vgl. Kapitel 3).

grade unterscheiden. Man spricht in diesem Fall von freien Fügungen, präferierten Kombinationen (Kollokationen) und festen bzw. fixen Fügungen.

Freie Fügungen kombinieren Wörter gemäß den grammatischen Regeln, etwa *schwarzer Pullover*.

Kollokationen zeigen eine Präferenz gegenüber anderen ebenfalls möglichen Ausdrücken, etwa *dickes Buch* gegenüber ?*breites Buch* oder **übergewichtiges Buch*.

Feste Fügungen sind so fixiert, dass keine alternativen Bildungen mehr möglich sind, vgl. *harte Droge* gegenüber **feste Droge*.

Ein bekanntes Beispiel für die unterschiedlichen kombinatorischen Möglichkeiten präferierter Fügungen ist *starker Tee* im Sinne von *intensiver Tee*, alternativ auch *kräftiger Tee*. Im Kontrast dazu kann die Neigung eines Rauchers zu intensivem Konsum mit *starker Raucher* bezeichnet werden, die Bildung **kräftiger Raucher* ist hier jedoch nicht analog verwendbar.

2.1.1 Wörter und lexikalische Einheiten

In Anbetracht dieser Tatsachen kommt man nicht umhin, sich Gedanken über die Definition des Begriffes „Wort“ zu machen. Analysiert man das Wort „Wort“ im Detail, stellt sich nach [A. M. Di Sciullo 1987, 1] heraus, dass damit bis zu vier verschiedene Konzepte bezeichnet werden können, und zwar das Wort als

1) Rahmenobjekt atomarer morphologischer Prozesse (im Wort finden morphologische Prozesse statt, d. h. Morpheme sind die Atome des Wortes)

2) Atomares Subjekt syntaktischer Prozesse auf Satzebene (Wörter sind die Atome der Syntax auf Satzebene)

3) Aufzulistende lexikalische Einheit (Listem) und als

4) Phonologische Einheit.

Bei der intuitiven Wahrnehmung von Wörtern handelt es sich meist um das vierte Konzept, gelegentlich auch um das zweite. Für die Diskussion im Rahmen dieser Arbeit ist jedoch Definition 3 zentral: das Wort als lexikalische Einheit, die in einem Lexikon verzeichnet werden muss. Damit wird zugleich klar, dass Fragen von

Morphologie und Syntax hier nicht im Vordergrund stehen: Entscheidend ist, dass alles, was nicht regelhaft abgeleitet werden kann, explizit im Lexikon kodiert werden muss - sowohl in computerlinguistischen Ressourcen als auch im (angenommenen) mentalen Lexikon.

> The lexical treatment of natural language, especially in the context of natural language processing requieres [sic] that the vocabulary of a given language is encoded in the form of electronic dictionaries and that it can be accessed in various kinds of applications.
>
> [F. Guenthner 2004, 1]

Das hier erwähnte Vokabular besteht zum einen aus allen einfachen Wörtern einer Sprache sowie den Regeln zum Ableiten von regelmäßig flektierbaren Wörtern. Zum anderen müssen aber auch alle „Sonderfälle“[6], d. h. nicht den üblichen Regeln unterworfene Einheiten, hier kodifiziert werden. Neben unregelmäßig flektierten einfachen Wörtern zählen dazu selbstverständlich auch diejenigen Wortgruppen, die gemeinsam etwas anderes bedeuten, als sie den Regeln entsprechend eigentlich müssten und die ich als MWEs oder „komplexe lexikalische Einheiten“ bezeichne. Ein vollständiges Lexikon müsste also enthalten:

- alle einfachen lexikalischen Einheiten
- alle komplexen lexikalischen Einheiten

Diese beiden Gruppen lassen sich mit dem Begriff „Listeme einer Sprache“ zusammenfassen. Regelmäßig gebildete flektierte Formen von einfachen und komplexen lexikalischen Einheiten sowie alle Phrasen, Satzteile und Sätze, die sich regulär bilden lassen, müssen nicht gelistet werden. Oder, auf den Punkt gebracht:

> As far as we can see, there is nothing more to say about them [idioms, d.V.] than (1) they are syntactic objects and (2) they are listed because of their failure to have a predictable property (usually their meaning).
>
> [A. M. Di Sciullo 1987, 5]

[6] Die Auffassung, dass es sich hier um „Sonderfälle“ oder „Ausnahmen“ handelt, ist ein gängiges Missverständnis, dem auch Anna Maria Di Sciullo and Edwin S. Williams in „On the Definition of Word“ von 1987 unterliegen. In Abschnitt 3.1.3 wird belegt, dass diese Bezeichnung schon rein quantitativ nicht gerechtfertigt ist.

Ein Beispiel: Die einfache lexikalische Einheit *schwarz* und die komplexe lexikalische Einheit *Schwarzes Loch* ähneln sich insofern, als dass beide nicht durch produktive Wortbildungsregeln erklärbar sind, sondern als Einheiten im Lexikon verzeichnet werden müssen. *Schwarz* ist im ersten Fall ein arbiträres, nicht motiviertes Zeichen, dessen Bedeutung schlicht gelernt werden muss. *Schwarzes Loch* dagegen besitzt eine konventionalisierte Bedeutung, die sich nicht allein aus *Schwarz + Loch* ableiten lässt. Auf MWEs bezogen heißt das: Selbst wenn alle Komponenten motiviert erscheinen und die Gesamtbedeutung transparent wirkt (etwa weil ein *Schwarzes Loch* tatsächlich schwarz erscheint), bleibt diese Motiviertheit bestenfalls plausibel, aber keinesfalls zwingend. Zum Vergleich: Der Ausdruck *schwarzes Hemd* lässt sich mit sehr hoher Wahrscheinlichkeit vollständig aus den Bedeutungen seiner Komponenten herleiten.

Die verschiedenen Phänomene natürlicher Sprachen, die unter den Begriff MWE fallen können, sind äußerst vielfältig. Die folgende Übersicht soll dies exemplarisch verdeutlichen, ohne den Anspruch einer systematischen Klassifikation zu erheben, auch die Terminologie wird an dieser Stelle noch einigermaßen naiv zur Illustration genutzt:

- **Idiome und stehende Wendungen** - semantisch nicht transparente, festgeprägte Ausdrücke: *ins Gras beißen*, *die Flinte ins Korn werfen.*
- **Kollokationen und feste Wortverbindungen** - lexikalisierte Kombinationen mit eingeschränkter Austauschbarkeit: *harte Drogen*, *Pfeffer und Salz.*
- **Terminologisierte Ausdrücke / Eigennamen** - komplexe Benennungen mit fester Referenz oder Fachspezialisierung: *Schwarzes Loch*, *Vereinte Nationen.*
- **Funktionsverbgefüge / Hilfsverbverbindungen** - analytische Konstruktionen mit hohem Konventionalisierungsgrad: *eine Rede halten*, *eingekauft haben.*
- **Weitere feste Strukturen** - etwa Sprichwörter (*Morgenstund hat Gold im Mund*), formelhafte Sprache (*meine Damen und Herren*) oder partikelartige Wendungen (*nach wie vor*).

2.1.2 Terminologie

Die Forschung um MWEs hat seit jeher mit einer bemerkenswerten terminologischen Vielfalt zu kämpfen. Friederike Schmidt kommentiert die hohe Zahl der vorgeschlagenen Bezeichnungen pointiert mit den Worten:

> [...] this number is almost as high as the number of linguists researching this subject.
>
> [F. Schmidt 2001, 7]

Damit benennt sie präzise den Kern des Problems: Die Vielzahl konkurrierender Termini erschwert nicht nur die Kommunikation innerhalb der Forschungsgemeinschaft, sondern verweist zugleich auf unterschiedliche theoretische Vorverständnisse. Terminologische Uneinheitlichkeit ist daher keineswegs eine Unannehmlichkeit, sondern spiegelt den fragmentierten Charakter des Forschungsfeldes insgesamt wider. Diese Beobachtung gilt nicht allein für den deutschsprachigen oder internationalen Diskurs, sondern zeigt sich in ähnlicher Weise auch in der hispanistischen Tradition. Besonders deutlich macht dies María Teresa Zurdo, wenn sie ausführt:

> Die Vielfalt und Heterogenität der spanischen Ausdrücke, die [...] in der einschlägigen Literatur verwendet werden, ist als unmittelbare Folge der Umfunktionalisierung der konservativen Terminologie zum einen und des Auftretens von neu geprägten Bezeichnungen zum anderen zu verstehen, welche das Aufkommen der Phraseologie als Teildisziplin der Linguistik mit sich bringt.
>
> [M. T. Zurdo 2007, 705]

Der Grund ist also (dies gilt für das Deutsche, das Spanische, das Englische und wohl auch für die meisten anderen europäischen Sprachen), dass das Phänomen der MWEs nicht von vorneherein als ein komplexes und entscheidendes Problem erkannt wurde. Vielmehr wurde es nach und nach und aus unterschiedlichen Forschungsrichtungen erfasst und beschrieben.

Erst mit dem Aufkommen der Phraseologie als eigener Forschungsrichtung begann die Suche nach einer gemeinsamen Terminologie. Einige Bezeichnungen lassen sich klar einer vorwissenschaftlichen Frühphase der Phraseologie zuordnen. Auch heute noch deutet die verwendete Terminologie häufig auf den jeweiligen disziplinären Zugang hin: aus literaturwissenschaftlicher Perspektive etwa *Redewendung*

oder *Leerformel*, aus der phraseologischen *Phrasem* oder *Phraseologismus*, aus der lexikographischen *Mehrwortlexem* oder *Phraseolexem*, aus der sprachhistorischen *erstarrte Wortfügung* oder *usuelle Einheit des Sprachschatzes*, aus der semantischen *Idiom*, aus der syntaktischen *mehrgliedrige lexikalische Einheit*, aus der korpuslinguistischen *Kollokation*.

Problematisch ist, dass die meisten Autoren ihre Terminologie mit einer fast selbstverständlichen Unbefangenheit verwenden, ohne die damit verbundenen theoretischen Implikationen zu reflektieren. Angesichts der Vielfalt an konkurrierenden Bezeichnungen halte ich eine argumentative Begründung der verwendeten Terminologie jedoch für notwendig.

Eine umfängliche Literatursuche zu Themen mit Bezug auf MWEs wird durch diese extreme Bandbreite von Ausdrücken jedoch maßgeblich erschwert: Wichtige Arbeiten finden sich mal unter dem Schlagwort „Phraseologie“, mal unter „Term“ und mal unter „Mehrworteinheit“. In Anhang C habe ich den Versuch unternommen, eine sehr umfangreiche Auflistung der mehr oder weniger gebräuchlichen Terminologie im Deutschen, Englischen und Spanischen zu erstellen. Jeder Begriff wird zusammen mit einem Beleg genannt. Wichtig ist, darauf hinzuweisen, dass die Begriffe in Anhang C durchaus nicht alle dasselbe bezeichnen, sondern häufig nur Teilmengen dessen, was der Begriff MWE in seiner weitesten Ausprägung umfasst. Dennoch handelt es sich bei dieser Liste um ein wichtiges Arbeitsmittel für die Recherche.

> Die zentrale Rolle bei der Benennung und Definition spielen m. E.: Mehrwortcharakter und Einheitsstatus. Der Mehrwortcharakter ist eine unabdingbare Prämisse, die verdeutlicht, daß [...] Sprachphänomene erfaßt werden sollen, die aus mindestens zwei Wörtern bestehen. Das Kriterium des Einheitsstatus betont die Wahrnehmung und Memorierbarkeit sprachlicher Phänomene als ein Ganzes, ohne sich in bezug auf die innere Stabilität dieser Einheiten bereits festzulegen und ohne in jedem Falle von einer Gesamtbedeutung, die nicht aus ihren Einzelbedeutungen erschließbar ist, auszugehen. Dabei werden dezidiert solche Erscheinungen ausgeschlossen, die nicht als Einheit gespeichert und reproduziert, sondern, wie die Kollokation, erst im Sprechakt aus einzelnen Teilen zusammengesetzt werden. [...] Es ist sinnvoll, diese Definition auch im Terminus aufzugreifen.
>
> [E. Donalies 1994, 344f]

Elke Donalies kam im Rahmen ihrer Argumentation im Übrigen zu dem Entschluss, den Ausdruck „Phrasem“ zu verwenden, gegen den ich mich jedoch entschieden habe. Dies ist darin begründet, dass der Begriff „Phrase“ in der CL je nach Kontext sowohl motivierte als auch unmotivierte Wortgruppen umfassen kann. In „Phrase-based Statistical Machine Translation“ etwa werden mit „Phrases“ beliebige nebeneinanderstehende Wortgruppen bezeichnet, gleichzeitig ist beispielsweise eine Nominalphrase streng definiert. Zusätzlich bezeichnet „Phrase“ auf Englisch mitunter einen ganzen Satz, ebenso ist „frase“ auf Spanisch ambig. Des Weiteren behandelt die Phraseologie nicht alle unter die weite Definition von MWE fallenden Phänomene, sondern hat ihren Fokus auf Redensarten und Redewendungen[7].

Im Weiteren verwende ich den Ausdruck Mehrworteinheit, abgekürzt MWE, als Oberbegriff. „Komplexe lexikalische Einheit“ setze ich synonym ein, vor allem dort, wo der Kontrast zur „einfachen lexikalischen Einheit“ sichtbar sein soll. Die Wahl folgt den Kriterien Mehrwortcharakter und Einheitsstatus aus [E. Donalies 1994] und schließt zugleich an die Abkürzung von „Multiword Expression“ sowie den Wortlaut von „Unidad Multipalabra“ an. Den Vorschlag von Kunze und Lemnitzer [C. Kunze 2007, 279], den schwer zu definierenden Terminus „Wort“ zu meiden und stattdessen „mehrgliedrige lexikalische Einheit“ zu bevorzugen, erkenne ich zwar an, entscheide mich jedoch dennoch aus Gründen der Leserführung und des internationalen Anschlusses für „Mehrworteinheit“. Auf dieser Grundlage verweise ich für die formale Arbeitsdefinition auf Abschnitt 2.1.3.

Der Begriff Idiom und davon abgeleitete Bezeichnungen erweisen sich als ungeeignet, da die Idiomatizität kein bestimmendes Kriterium von MWEs darstellt (vgl. [E. Donalies 1994, 345] sowie die Diskussion der Kriterien weiter unten).

2.1.3 Arbeitsdefinition

Es gibt fast so viele Definitionen von MWEs wie es Bezeichnungen dafür gibt. Die wohl meistzitierte Definition ist von Ivan A. Sag, Timothy Baldwin et al. aus dem einflussreichen „A Pain in the Neck“ Artikel von 2002 und sehr knapp gehalten:

> We define multiword expressions (MWEs) very roughly as “idiosyncratic interpretations that cross word boundaries (or spaces)”.
>
> [I. A. Sag 2002, 2]

[7] Dies trifft zumindest auf die Phraseologie im engeren Sinne zu. Soweit eine Phraseologie im weiteren Sinne im Gespräch ist, bezieht sich diese meines Wissens nach zumindest von ihrem Anspruch her auf alle hier beschriebenen Phänomene.

Diese Definition ist eine gute instruierende Arbeitsgrundlage, aber nicht formal genügend. Basierend auf den genannten Kriterien definiere ich MWE für diese Arbeit wie folgt:

Definition: Mehrworteinheit (MWE) Treten mehrere lexikalische Einheiten - einfach oder rekursiv-komplex (verschachtelt) - gemeinsam auf (auch diskontinuierlich oder satzförmig) und werden dabei im Sprachgebrauch oder in der Verarbeitung als zusammengehörige Einheit behandelt (nicht zwingend als syntaktische Konstituente), die sich auf mindestens einer Ebene (lexikalisch, syntaktisch, distributionell, transformationell, semantisch, pragmatisch oder interlingual) idiomatisch verhält, so handelt es sich um eine komplexe lexikalische Einheit („Mehrworteinheit"). Solche Einheiten sind häufig lexikalisiert, müssen aber in jedem Fall als Listeme in elektronischen Lexika erfasst werden.[8]

Im Folgenden präzisiere ich die in der Definition genannten Dimensionen als Arbeitskriterien.

2.2 Kriterien für MWE

Im Rahmen der Definition habe ich bereits einige Kriterien für MWEs angesprochen. Diese sollen im folgenden Kapitel ausführlicher dokumentiert werden. Zum Teil bestehen in der Literatur, ganz der Tradition der terminologischen Vielfalt folgend, voneinander abweichende Bezeichnungen für dieselben oder sehr ähnliche Sachverhalte, welche ich daher ebenfalls anführe. Die angegebenen Quellen sind lediglich Belege der Verwendung, keine systematische Bestimmung der Erstverwendung.

Es gibt eine Reihe von Kriterien, anhand denen sich reguläre Bildungen wie *schwarzer Pullover* von MWEs wie *Schwarzes Loch*[9] unterscheiden lassen. Sie betreffen alle linguistischen Ebenen, von der Syntax über morphosyntaktische und semantische Merkmale bis hin zum distributionellen und pragmatischen Verhalten. Jedoch ist keines der Einzelkriterien für sich genommen notwendig oder hinreichend, um eine MWE formal zu bestimmen - außerdem sind die Kriterien nicht alle für die (teil-)automatisierte Extraktion aus Korpora anwendbar. Die Kriterien sind im Einzelnen:

[8] Kurzfassung: „Treten mehrere lexikalische Einheiten gemeinsam auf und werden als Einheit behandelt, die sich mindestens auf einer Ebene idiomatisch verhält, handelt es sich um eine Mehrworteinheit (und damit um ein Listem)."

[9] Zur Diskussion genau dieses Beispiels siehe die Klassifikation von Guenthner und Blanco, Abschnitt 2.3.2.

2.2.1 Polylexikalität

Bezeichnungen in der Literatur: Polylexikalität ([M. T. Zurdo 2007]), Pluriverbalität ([B. A. Model 2010]), Mehrwortcharakter ([E. Donalies 1994])

Eines von zwei bestimmenden Kriterien für MWE ist deren Polylexikalität, d. h. die Tatsache, dass es sich „um Gebilde aus zwei oder mehr Wörtern handelt“ ([E. Donalies 1994, 336]). Dies ist in dem Fall unstreitig, in dem davon ausgegangen wird, dass Leerzeichen und Interpunktionszeichen wie Bindestriche als Wortgrenzen anerkannt werden (also auch bei *kaufte ein* vs. *einkaufen*) und gleichzeitig zuerkannt wird, dass mitunter kein Gebrauch davon gemacht wird (also dass spanisch *dámelo* als *da me lo* betrachtet wird). Gleichzeitig werden Affixe und Derivationsmorpheme nicht als „Wörter“ betrachtet, sofern diese nicht durch Leerzeichen oder Interpunktionszeichen von den zu verändernden Wörtern getrennt sind. Die einzelnen Einheiten, aus denen MWE zusammengesetzt sind, werden als Elemente oder Konstituenten bezeichnet.

2.2.2 Einheitsstatus

Bezeichnungen in der Literatur: Einheitsstatus ([E. Donalies 1994]), Satzgliedwertigkeit ([M. T. Zurdo 2007])
Mit der Bezeichnung „Einheitsstatus“ wird einerseits auf die Eigenschaft zahlreicher MWE verwiesen, sich trotz ihrer Polylexikalität syntaktisch wie ein einzelnes Wort zu verhalten, ([I. A. Sag 2002]) bezeichnet sie als „words with spaces“. Diese Deutung von „Einheitsstatus“ wird auch von [E. Donalies 1994] angenommen und findet ihre Entsprechung in der Bezeichnung „Satzgliedwertigkeit“, was bedeutet, dass die MWE die Funktion eines Satzgliedes einnimmt.

Es gibt jedoch auch Typen von MWEs, die über innere Strukturen verfügen und somit nicht wie „ein Wort“ behandelt werden dürfen.

Entsprechend ist die Definition des Begriffs „Einheitsstatus“ zu erweitern, um ein entscheidendes Kriterium für alle MWEs zu formulieren. Mehrere, mitunter über den Satz hinaus verteilte Einzelwörter werden als zusammengehörig behandelt, d. h. auch diskontinuierliche Wortfolgen werden als eine Einheit aufgefasst, ohne dass damit bereits etwas über ihr syntaktisches Verhalten ausgesagt wäre. Mit anderen Worten: Sie bilden eine Einheit, aber nicht notwendig eine syntaktische.

2.2.3 Idiomatizität

Bei Idiomatizität handelt es sich um ein wichtiges Kriterium für viele Typen von MWEs, das unterschiedlich weit gefasst werden kann, von einer rein semantischen Perspektive bis zu jeglicher Abweichung von erwartbarem Verhalten. In dieser Arbeit wird der weite Begriff zugrunde gelegt, und die unterschiedlichen Kategorien werden im Anschluss beschrieben.

[M. Almela Sánchez 2006] definiert Idiomatizität als „[...] solidarities between meanings and multi-word patterning", d. h. der Fokus liegt hier genau wie bei den anderen in der Literatur verwendeten Bezeichnungen klar auf einer semantischen Idiomatizität. Idiomatizität kann jedoch auch ganz allgemein ein vom Vorhersagbaren abweichendes Verhalten von lexikalischen Einheiten bedeuten:

> In the context of MWEs, *idiomaticity* refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical level. A given MWE is often idiomatic at multiple levels.
>
> [T. Baldwin 2010, 4]

Eine entscheidende Frage ist nun, ob Idiomatizität (wie etwa [T. Baldwin 2010] schreibt) tatsächlich eine notwendige Bedingung für MWEs darstellt, oder ob sie es nicht ist (wie [E. Donalies 1994] argumentiert).

Wird die rein semantische Idiomatizität herangezogen, ist Elke Donalies sicherlich im Recht: Es gibt zahlreiche MWEs, die semantisch sehr wohl transparent aber dennoch zweifelsfrei als komplexe lexikalische Einheiten wahrzunehmen sind (z. B. Wendungen wie *mehr Glück als Verstand haben* vgl. [E. Donalies 1994, Anm. 41]).

In der weiter gefassten Definition von Idiomatizität hingegen, wie sie Timothy Baldwin und Su Nam Kim vertreten (jedes abweichende Verhalten ist als idiomatisch zu bezeichnen), ist tatsächlich davon auszugehen, dass die überwältigende Mehrheit der MWEs (wenn nicht sogar alle) in wenigstens einer Form idiomatisch sind.

Insofern findet dieses Kriterium auch in der von mir verwendeten Definition neben dem Mehrwortcharakter und dem Einheitsstatus Anwendung.

Viele idiomatische Phänomene sind graduell ausgeprägt und schwer eindeutig zu bestimmen, schon gar nicht maschinell.

> Paradoxerweise lässt sich der Begriff der Idiomatizität schwer eindeutig definieren, und der Grad seiner Intensität objektiv messen. Bei der Bestimmung des Wesens der Idiomatizität sollte man von der allgemeinen

> Vorstellung ausgehen, dass die Idiomatizität sprachlicher Strukturen eine gewisse Irregularität der betreffenden Einheiten der Sprache voraussetzt.
>
> [D. Dobrovol'skij 2023, 51]

Die folgende Übersicht dient daher insbesondere der ausführlichen Begründung der obigen Definition und nicht zwangsläufig der Identifikation. Vielmehr sind die folgenden Kriterien eine hilfreiche Handreichung für die notwendige händische Kuration von MWEs wie `LEMoN.dic`.

Im Folgenden sollen einige Aspekte genauer betrachtet werden, die in der Literatur häufig als eigene Charakteristika herangezogen werden, meiner Ansicht nach aber allesamt verschiedene Aspekte von Idiomatizität darstellen:

2.2.4 Formale Kriterien

Syntaktische Idiomatizität: Syntaktische Idiomatizität (oder „extragrammatical idioms“ bei [C. J. Fillmore 1988]) beschreibt ein Phänomen, das sich sowohl in der internen Struktur einer MWE als auch in ihrem äußeren Satzverhalten manifestieren kann (vgl. [T. Baldwin 2010, 10], [H.-R. Chae 2015]). Während interne syntaktische Besonderheiten häufig darin bestehen, dass die Bauweise nicht regulär aus den Komponenten abgeleitet werden kann (z. B. *de vez en cuando*, *by and large*), betrifft externe Idiomatizität die Art und Weise, wie eine MWE sich im Satz verhält und in welchen syntaktischen Positionen sie auftreten kann (z. B. *kingdom come*, formal eine fossilierte `N+V`-Struktur, distributionell jedoch meist adverbial verwendet, etwa in *send you to kingdom come*).

Syntaktische und semantische Idiomatizität können in wechselseitiger Beziehung stehen:

> Syntactic idiomaticity is related to semantic idiomaticity in that an idiosyncratic syntax can preclude the possibility of a literal, i.e. non-idiomatic, interpretation of a string.
>
> [J. Y. Findlay 2019, 23]

In der Literatur werden gelegentlich unter dieser Bezeichnung nicht nur syntaktische sondern auch morphosyntaktische oder rein morphologische Phänomene beschrieben, insofern ist syntaktische Idiomatizität auch als Kontinuum zu begreifen:

> Although idioms are generally assumed to be non-compositional and, hence, non-flexible, it has been well attested that they are not fixed expressions formally. Even one of the most fixed idioms like *kick the bucket* shows morphological flexibility in the behavior of the verb *kick*. Many other idioms show some degree of syntactic flexibility with reference to various types of syntactic behavior.
>
> [H.-R. Chae 2015, 46]

Lexikalische Idiomatizität: Von lexikalischer Idiomatizität spricht man, wenn sich einzelne oder alle Komponenten einer MWE ungewöhnlich hinsichtlich ihres lexikalischen Gebrauchs verhalten. Auch hierbei handelt es sich um ein graduelles Phänomen: Der Grad reicht von selten belegten Lexemen bis zu Wörtern, die ausschließlich innerhalb spezifischer MWEs vorkommen, sogenannte Unikale. Typische Beispiele umfassen Wendungen wie *a troche y moche* oder *de pe a pa*. Solche Bestandteile werden in der Literatur auch als „cranberry words“ bezeichnet. Lexikalische, syntaktische und semantische Idiomatizität bedingen einander häufig:

> Since such phrases are not part of the domestic vocabulary, lexical idiomaticity usually entails semantic idiomaticity, since the meaning of the expression is then usually not deduced from the meaning of the parts [...]. For the same reason, lexical idiomaticity usually also entails syntactic idiomaticity.
>
> [G. I. Lyse 2012, 82]

2.2.5 Kontextuale Kriterien

Distributionelle Idiomatizität: Mit distributioneller Idiomatizität wird auf die Eigenschaft einer großen Gruppe von MWEs verwiesen, die zwar nicht opak sind, aber „bevorzugte Versprachlichungen der jeweiligen Sachverhalte darstellen“ ([B. Pöll 2002]84). Diese MWEs werden oft als Kollokationen (oder Kookkurrenzen) bezeichnet, jedoch kann der Begriff der Kollokation je nach Deutung auch sämtliche Wortgruppen behandeln, die häufig in Gegenwart voneinander auftreten, auch wenn diese zufällig sind bzw. offensichtlich keine Einheiten bilden.

Der Grad der distributionellen Idiomatizität reicht von zufälligen Verbindungen (*dir einen*) über bevorzugte Versprachlichungen (*tiefe Trauer*) bis zu exklusiven Verbindungen (*gang und gäbe*). Ab einem bestimmten Grad von Idiomatizität ist die

Grenze zur Lexikalisierung überschritten, d. h. diese Kombination wird vom Sprecher nicht länger als frei wahrgenommen, sondern als feststehend (siehe Abschnitt: 2.2.6). Die Exklusivität bemisst sich jedoch nicht ausschließlich daran, ob die jeweiligen Elemente einer MWE gemeinsam gehäuft auftreten und ob sie in jeweils anderen gültigen Kombinationen selten bis überhaupt nicht auftreten. In vielen Fällen sind unspezifische Bildungen (z. B. *Angst machen*) wesentlich häufiger in Texten anzutreffen als spezifische, d. h. „in ihrer Kombinatorik eingeschränkte“ ([B. Pöll 2002]84) (*Angst einflößen*).

Distributionelle Idiomatizität ist ein wichtiges Charakteristikum vieler MWEs, aber – genau wie semantische Idiomatizität – kein eindeutiger Marker. In der Literatur wird dieses Feld unter unterschiedlichen, nicht deckungsgleichen Bezeichnungen diskutiert, darunter „lexical distributional similarity“ [J. E. Weeds 2003], „statistical idiomaticity“ [T. Baldwin 2010], „statistische Idiosynkrasie“ [I. A. Sag 2002], „Kollokation“ [J. R. Firth 1957] und „Kookkurrenz“ [B. Daille 1994].

2.2.6 Interpretative Kriterien

Semantische Idiomatizität: Diese Form der Idiomatizität ist bei manchen Autoren die prototypische oder einzige Form, insofern wird gelegentlich mit „Idiomatizität“ ausschließlich semantische Idiomatizität gemeint ([H. Burger 2010], [M. T. Zurdo 2007]). Sie wird auch als (Nicht-)Kompositionalität ([T. Baldwin 2010]), Konventionalität ([G. Nunberg 1994]) oder semantische Opazität ([G. Gross 1996]) bezeichnet. Die Bedeutung einer MWE ergibt sich häufig nicht (nur) aus den Bedeutungen ihrer Bestandteile, sie ist in diesem Fall also nichtkompositionell. Man spricht in diesem Zusammenhang auch von opaker Semantik (im Gegensatz zu transparenter Semantik) oder Opazität.

Diese wiederum lässt sich nach [H. Burger 1982] grob in die Gruppen „Motiviert“, „Teilmotiviert“ und „Unmotiviert“ oder umgekehrt in „Vollidiomatisch“, „Teilidiomatisch“und „Transparent“ einteilen. Das Problem an dieser Definition ist die im Einzelfall oft schwierige Festlegung darauf, ob und wie sehr eine lexikalische Einheit nun motiviert ist. Dazu muss einerseits definiert werden, welche Lesarten einer Konstituente nun regulär sind und welche übertragen. Des Weiteren kommt hinzu, dass viele Sprecher auch in treffenden Metaphern und in klanglichen Varianten durchaus Motiviertheit erkennen. Im Übrigen können auch einfache lexikalische Einheiten idiomatisch sein, d. h. etwa im übertragenen Sinne verwendet werden.

Eine weitere Unterscheidung betrifft „endozentrische“ und „exozentrische“ MWEs.

Endozentrische MWEs erben (zumindest teilweise) Bedeutung, Kategorie oder syntaktisches Verhalten von einem Kopf innerhalb der Einheit, exozentrische hingegen nicht. So bezeichnet *hombre-rana* tatsächlich einen Mann und verhält sich kategorial und syntaktisch wie *hombre* (Plural *hombres-rana* usw.). Demgegenüber liegt bei *piel roja* keine „spezifische Form von Haut“ ([B. Pöll 2002, 35]) vor, sondern eine rassistische Bezeichnung für amerikanische Ureinwohner. Die Einheit verhält sich semantisch und distributionell wie ein menschenbezeichnendes Nomen und nicht wie *piel*.

Lexikalisierung: Lexikalisierung liegt vor, wenn eine komplexe lexikalische Einheit teilweise oder vollständig demotiviert ist, wenn sie konventionalisiert und damit im mentalen bzw. sozialen Lexikon verankert ist oder wenn sie eine erhöhte Festigkeit gegenüber Variation zeigt. Je nach Ausprägung kann dies mit Idiomatizität einhergehen.

Der Abstand zwischen der transparenten Bedeutung eines kompositionell gebildeten Ausdrucks und seiner tatsächlichen Bedeutung kann als Gradmesser für die semantische Idiomatizität des jeweiligen Ausdrucks dienen. Je höher der Grad der Idiomatizität, desto wahrscheinlicher ist, dass eine komplexe Einheit lexikalisch im Wortschatz verankert ist.

Autoren wie etwa [B. Pöll 2002] gehen davon aus, dass eine MWE beim Prozess der Lexikalisierung zwangsläufig eine semantische Idiomatisierung durchläuft. Allerdings zeigt sich an zweifellos lexikalisierten, aber semantisch transparenten Einheiten wie *zumo de naranja*, dass Lexikalisierung nicht zwingend mit semantischer Idiomatisierung einhergeht, vgl. für das Spanische auch [C. Buenafuentes de la Mata 2007]. Versteht man Idiomatizität jedoch breiter - etwa auch als distributionelle oder formale Beschränkung -, lässt sich dieser Widerspruch auflösen.

Im Gegenzug ist eine Idiomatisierung ohne eine Lexikalisierung nicht denkbar, da die Sprecher sonst nicht in der Lage wären, den entsprechenden Ausdruck zu verwenden. Alternative Bezeichnungen für Lexikalisierung sind Institutionalisierung ([T. Baldwin 2010]), Konventionalisierung ([T. Baldwin 2010]), Reproduzierbarkeit ([B. Pöll 2002]), Geläufigkeit ([M. T. Zurdo 2007]), Verknöcherung ([J. Grimm 1854]) oder Sprachüblichkeit ([E. Donalies 1994])

Transformationelle Idiomatizität: Im Gegensatz zu den in der Regel flexibleren freien Syntagmen sind viele MWEs einem Phänomen unterworfen, das unter an-

derem mit Ausdrücken wie Stabilität, Erstarrtheit oder Fixierung bezeichnet wird. Damit ist gemeint, dass MWE häufig nicht die gleichen Freiheiten wie freie Verbindungen genießen, sondern gewissen Beschränkungen unterliegen. Diese Beschränkungen können laut [B. Pöll 2002]

- lexikalisch-semantischer Natur sein (d. h. Elemente sind nicht durch Synonyme zu ersetzen, etwa *canela fina* vs. **canela delicada*),
- sich auf die Abfolge der Komponenten beziehen (*al fin y al cabo* vs. **al cabo y al fin*),
- den Verwendungskontext determinieren (pragmatische Fixierung, etwa bei Grußformeln)
- oder hinsichtlich ihrer Transformierbarkeit gelten (transformationelle Defektivität, siehe unten).

Es lassen sich dabei verschiedene Grade von Fixiertheit ausmachen: Zum Einen kann man unterscheiden, wie viele Elemente einer MWE betroffen sind (nur eins, einige oder alle) und wie stark sie eingeschränkt sind (einige MWEs sind überhaupt nicht betroffen, andere sind vollkommen erstarrt).

Für nominale MWEs hat Friederike Schmidt die folgenden Transformationen als oft nicht möglich aufgestellt:

a) non-modifiability: *a very black hole

b) they cannot be pronominalized: *the hole which is black

c) their parts cannot be substituted by synonyms: *black orifice

d) coordination is not possible: *a black and deep hole

e) parts cannot be deleted: *this hole...

f) adjectives cannot be nominalized: *the blackness of the hole

[F. Schmidt 2001, 5]

Es lässt sich also daraus ableiten, dass MWEs ein irreguläres Verhalten bezüglich ihrer Transformierbarkeit aufweisen, denn keine der oben genannten Regeln lässt sich auf alle MWEs anwenden. Vielmehr lassen sich zwei einander entgegengesetzte Extreme ausmachen: die freie Verbindung auf der einen Seite und die vollkommen erstarrte Verbindung auf der anderen, d. h. keine Konstituente lässt irgendeine

Form von Variation zu, dies ist jedoch sehr selten (z. B. *klipp und klar*). Die meisten MWEs finden sich auf den Punkten zwischen diesen beiden Extremen. Weitere Ausführungen zu Art, Grad und Reichweite von Fixierungen finden sich etwa bei [G. Gross 1996].

Es ist ebenfalls ein Charakteristikum von erstarrten MWE, dass sie in Texten aus bestimmten Domänen (insbesondere in journalistischen Texten) eben doch einer Form von Variation ausgesetzt sind, und zwar dem Wortspiel. Die Abweichung von der Erwartung wird hier als Mittel zur Erregung von Aufmerksamkeit des Lesers eingesetzt (vgl. [G. Gross 1996, 20]).

Am Grad der Erstarrtheit stellt sich auch die Frage nach der syntaktischen Funktion einer MWE im Satz. Viele MWEs verhalten sich wie syntaktische Einheiten: Sie lassen sich wie einfache Substantive in eine Nominalphrase einbetten, als Nominalphrase verwenden und im Satz verschieben. Andere teilen die distributionellen Eigenschaften entsprechender Substantive oder solcher lexikalischer Einheiten derselben semantischen Klasse. Gleichwohl unterscheidet sie ihr Mehrwortcharakter deutlich vom Verhalten einfacher Wörter. Weitere Bezeichnungen in der Literatur sind „Inflexibilität“ ([G. Nunberg 1994]), „Festigkeit“ ([M. T. Zurdo 2007]), „Erstarrtheit und Stabilität“ ([E. Donalies 1994]), „Fixiertheit“ ([C. Kunze 2007]) und „Blockierung transformationaler Eigenschaften“ ([G. Gross 1996]).

Pragmatische Idiomatizität: Unter pragmatischer (auch: funktionaler) Idiomatizität versteht man Fälle, in denen eine MWE ihre spezifische idiomatische Bedeutung erst durch den Gebrauch in bestimmten kommunikativen oder situativen Kontexten entfaltet, und die von Grußformeln wie *Buenos días* über rituelle Wünsche wie *Hals- und Beinbruch* bis zu konventionalisierten Ausrufen wie *¡Todos a cubierta!* reichen.

Entscheidend ist hier nicht primär die innere Struktur oder semantische Opazität, sondern die fixierte Verwendung im Diskurs:

> [...] a speaker may use an idiom rather than the non-idiomatic, but propositionally equivalent content, when his/her communicational message includes information that is related to the reported context [...]
>
> [S. Alghazo 2022, 427]

Interlinguale Idiomatizität: Viele MWEs beschreiben Konzepte, die auch in anderen Sprachen mittels MWEs ausgedrückt werden. Diese sind gelegentlich lexi-

kalisch und strukturell analog (z. B. *black hole* vs. *agujero negro*), oft jedoch finden sich Abweichungen wie *Vom Regen in die Traufe* vs. *de la sartén al fuego*. Diese können von geringfügigen Elementen bis zu syntaktisch und/oder lexikalisch vollständig unterschiedlichen Konstruktionen reichen. Andere MWEs hingegen verfügen in anderen Sprachen über entweder keine direkten Übersetzungen und müssen gebildet werden oder aber sie verfügen über eine einfache lexikalische Einheit, die sie übersetzt (für Beispiele siehe [T. Baldwin 2010]). Da sich beide Phänomene in alignierten parallelen Korpora gut identifizieren lassen, sind diese gute Werkzeuge zur Identifikation von MWEs.

Auch bei der Entscheidung, ob es sich bei einer konkreten Wortgruppe um eine MWE handelt, ist es zuweilen hilfreich, theoretisch anzunehmen, ob das von ihr beschriebene Konzept in einer anderen Sprache auch mit einer einfachen lexikalischen Einheit übersetzt werden könnte. Diese Form der Idiomatizität wird auch als „Inter- oder Crosslinguale Variation“ ([T. Baldwin 2010]) oder „Translatorische Idiomatizität“ ([J. Monti 2011]) bezeichnet.

Vor diesem Hintergrund werden im nächsten Abschnitt relevante Klassifikationstraditionen von MWEs vorgestellt.

2.3 Klassifikation von MWEs

Die Klassifizierung von MWEs ist aufgrund der Heterogenität der unter diesen Begriff fallenden Phänomene ein zentraler Schritt zu deren Verständnis und Verarbeitung. In der Literatur finden sich zahlreiche Ansätze, die auf sehr unterschiedlichen theoretischen Annahmen beruhen und von wenigen bis zu über siebenhundert Kategorien[10] reichen. Zwar geht [M. T. Zurdo 2007] davon aus, in der Forschung herrsche

> [...] weitgehende Übereinstimmung auf dem Gebiet der wissenschaftlich geltenden Klassifikationskriterien [...]. Die Abweichungen, welche die einzelnen Klassifikationsansätze voneinander unterscheiden, beruhen eher auf sprachwissenschaftlich-theoretischen Ausrichtungen oder praktischen Zielsetzungen als auf divergierenden Meinungen bezüglich der Einordnungskriterien selbst.
>
> [M. T. Zurdo 2007, 706]

[10] Vgl. [M. Mathieu-Colas 1996], ausführlicher im Abschnitt zur Gross-Schule 2.3.2 und später in der Beschreibung der LEMoN-Grammatiken im Kapitel 9.

Diesem Urteil kann ich mich in Bezug auf die im Folgenden vorzustellenden Klassifizierungsansätze jedoch nicht anschließen. Die Ansätze unterscheiden sich durchaus stärker voneinander, als dass man von einer Übereinstimmung in den Kriterien und einer abweichenden Auswahl der für die eigene Arbeit relevanten ausgehen könnte. Vielmehr scheint nach wie vor Uneinigkeit über das zu untersuchende Phänomen an sich zu bestehen. Auch [C. Buenafuentes de la Mata 2021] sagt mit Blick auf „Compounds “ sehr richtig:

> If the twofold task of defining and delimiting the compound word has occasioned keen discussion in morphological theory [...], establishing a typology of compound words likewise is generally thought to be a highly controversial issue. Thus, it is possible to claim that up to now, there is not yet a universally well-established classification of compounds which has been jointly agreed upon regardless of the theoretical focus applied.
>
> [C. Buenafuentes de la Mata 2021, 285f]

Ähnlich wie die Terminologie ist auch die Klassifikation von MWEs breit und heterogen. Für die vorliegende Untersuchung stütze ich mich auf zwei Perspektiven: eine allgemeine, auf die Computerlinguistik ausgerichtete Sicht und die Sicht der Gross-Schule. MWE-Klassifikationen mit explizitem Fokus auf das Spanische berücksichtige ich an dieser Stelle nicht, verweise aber auf [C. Parra Escartín 2018] und [C. Buenafuentes de la Mata 2021]. Stattdessen fokussiere ich in Kapitel 3.2 explizit auf Taxonomien spanischer nominaler MWEs.

2.3.1 Allgemein (computer-)linguistische Taxonomien

Ivan Sag et al. (2002): Die von [I. A. Sag 2002] vorgestellte Klassifikation gilt als eine der einflussreichsten Einteilungen von MWEs. Der zugrunde liegende Aufsatz „Multiword Expressions: A Pain in the Neck for NLP“ spielte eine wichtige Rolle bei der Popularisierung des Themas in der Sprachverarbeitung, die zu dieser Zeit eine neue Blüte erlebte. Die eher problembeschreibende Kategorisierung knüpft an [L. Bauer 1983] und [G. Nunberg 1994] an und hat einen klaren morphosyntaktischen Fokus. Im Kern unterscheidet der Ansatz zwei Hauptklassen, die erste gliedert sich in drei Unterklassen, vgl. Abbildung 1.

Lexicalized Phrases beinhalten eine wenigstens teilweise Idiomatizität oder wenigstens Wörter, die nie isoliert auftreten. Sie lassen sich anhand der Fixiertheit, der sie unterliegen, in die folgenden Unterklassen einteilen:

Fixed Expressions Es handelt sich hierbei um vollständig lexikalisierte und unveränderliche Ausdrücke, das heißt sie können wie „Words with spaces“ behandelt werden, also als Einheit. Neben feststehenden Redensarten (*by and large*) betrifft dies auch generell auflösbare Verbindungen, die allerdings aufgrund ihrer fremdsprachlichen Herkunft für den größten Teil der Sprecher opak bleiben (*ad hoc*).

Semi-Fixed Expressions Die MWEs in dieser Kategorie unterliegen zwar ebenfalls strengen syntaktischen Regeln, sind jedoch nicht ganz so unveränderlich wie Fixed Expressions, sondern ermöglichen diverse Variationen wie die Verwendung reflexiver Formen. Hierzu zählen nichtzerlegbare Idiome, Eigennamen und Komposita.

Syntactically Flexible Expressions Im Unterschied zu den Semi-Fixed Expressions lassen Syntactically Flexible Expressions nicht nur Flexion und Insertion als Variationen zu, sondern erlauben auch die Umstellung ihrer Elemente. Dies beinhaltet unter anderem Verb-Partikel-Konstruktionen, zerlegbare Komposita sowie Stützverbkonstruktionen.

Institutionalized Phrases werden zwar regulär gebildet und sind transparent, treten jedoch statistisch signifikant gehäuft gemeinsam auf.

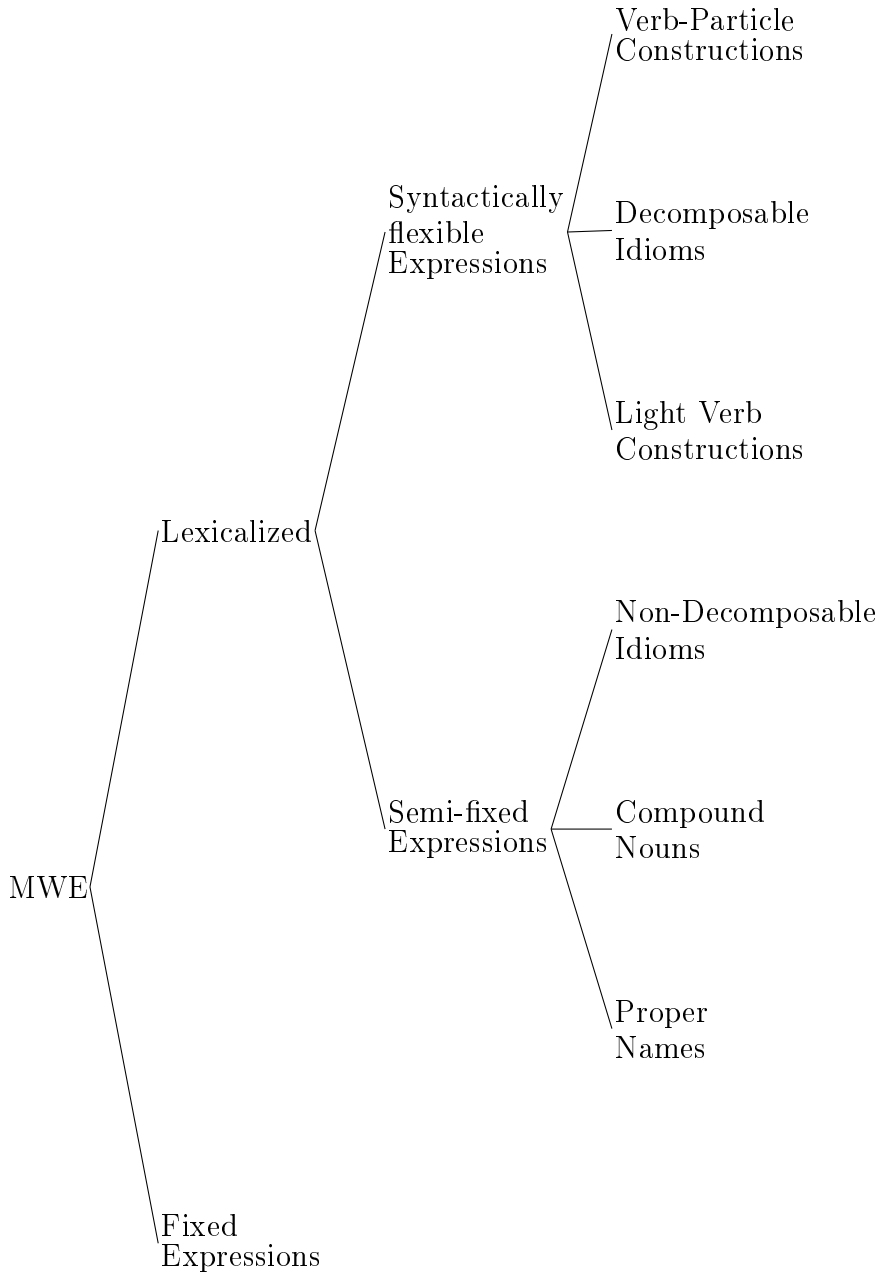

Abbildung 1: Klassifikation nach Sag et al. (2002)

Claudia Kunze und Lothar Lemnitzer (2007): In ihrer 2007 erschienenen „Einführung in die Computerlexikographie“ ([C. Kunze 2007]) stellen Claudia Kunze und Lothar Lemnitzer eine klare Arbeitsklassifikation mit sechs Kategorien vor. Die *Phraseme* entsprechen dabei den hier behandelten MWEs. Anstelle einer feinen Untergliederung der Phraseme betonen sie den Bedarf an weiterer korpusge-

stützter Empirie. Als hilfreichen Ausgangspunkt nennen sie die Unterscheidung bei [G. Nunberg 1994] in „idiomatisch kombinierte Ausdrücke“ und „feste Fügungen“. Die Taxonomie ist einschließlich dieser vorläufigen Unterscheidung in Abbildung 2 abgebildet.

> Diese Gruppierung in lediglich zwei Klassen ist aber noch zu grob für unsere Zwecke. Weitere, empirische und korpusgestützte, Untersuchungen sind notwendig, um herauszufinden, ob eine strikte Subklassifizierung von Phrasemen möglich ist. In der Zwischenzeit müssen wir Phraseme lexikalisch behandeln, d. h. die Beschränkungen für jedes einzelne Phrasem bestimmen.
>
> [C. Kunze 2007, 304]

Die Kategorien umfassen im Einzelnen

Phraseme Entsprechen den englischen *Idioms* und sind semantisch opak.

Kollokationen Zeichnen sich vor allem durch meist transparente Bindungen aus, die theoretisch synonyme lexikalische Kombinationen ausschließen oder markieren.

Mehrgliedrige Komposita Die vornehme Form der Kompositabildung in den romanischen Sprachen und auf Englisch.

Phrasale Verben und Partikelverben Verben, die durch Hinzufügung eines Funktionswortes oder Adverbs ihre Bedeutung verändern. Im Deutschen werden diese im Unterschied zu anderen Sprachen typischerweise in einer einfachen lexikalischen Einheit realisiert.

Mehrgliedrige Funktionswörter Nicht zwingend kontinuierliche Folgen von Funktionswörtern oder Adverbien.

Funktionsverbgefüge und Nominalisierungsverbgefüge Bestehen aus einem bedeutungsschwachen Verb und einem bedeutungstragenden Nomen bzw. einer Nominalphrase (NP).

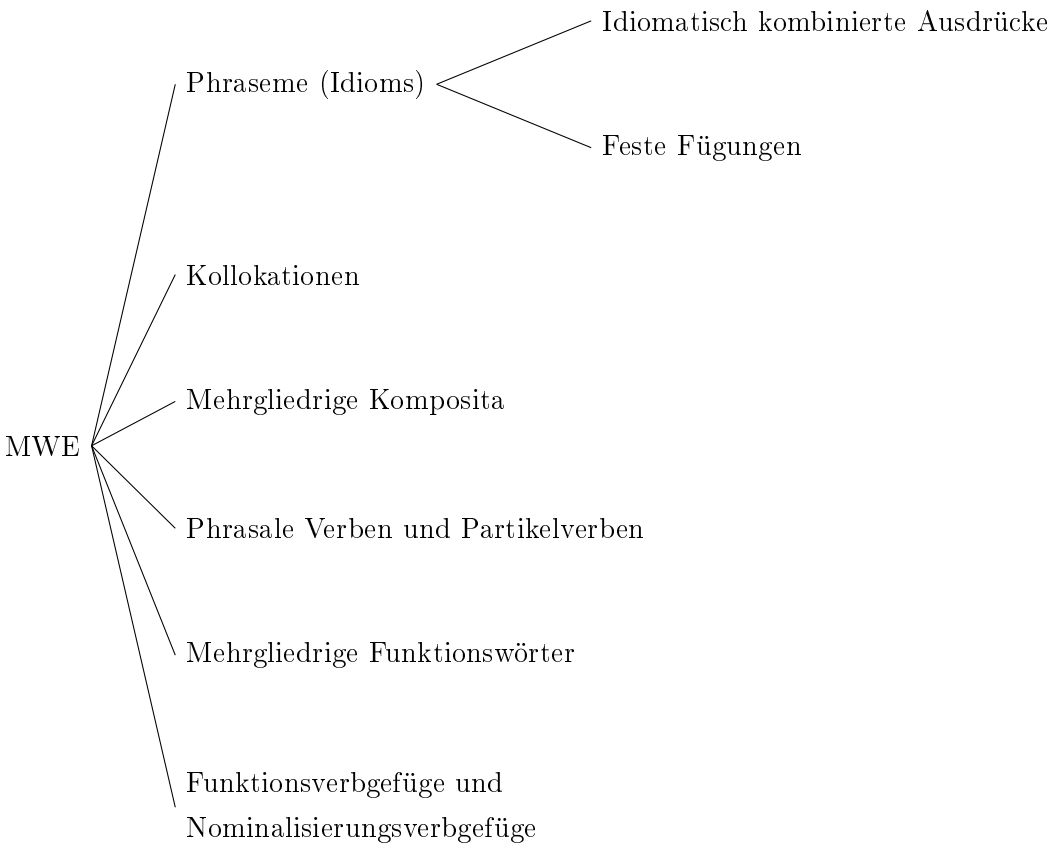

Abbildung 2: Klassifikation nach Kunze und Lemnitzer (2007)

2.3.2 Aus der Gross-Schule hervorgegangene Taxonomien

Bei der Betrachtung der Taxonomien, die aus der Gross-Schule hervorgegangen sind, muss an erster Stelle diejenige von Begründer Maurice Gross selbst stehen. Diese folgt nach einer kurzen Einführung zum wissenschaftlichen Kontext: Maurice Gross knüpfte in den 1960er-Jahren an die distributionalistische Tradition von Zellig S. Harris an und übertrug deren Konzepte auf lexikalische Inventare. Daraus entwickelte er die Theorie der Lexikon-Grammatik, die Lexikon und Grammatik nicht als getrennte, sondern als eng verflochtene Systeme versteht.[11]

Das Kernelement sind sogenannte „elementare Sätze“, also Sätze mit prototypischem Charakter und mit argumentativen Leerstellen, deren Eigenschaften in komplexen Tabellen systematisch dokumentiert werden. In diese Tabellen lassen sich sowohl einfache als auch komplexe lexikalische Einheiten eintragen, wodurch MWEs

[11] Für eine ausführlichere Darstellung des wissenschaftsgeschichtlichen Kontexts vgl. Abschnitt 2.5.

explizit und konsistent erfasst werden. Damit schuf Gross ein Verfahren, das die bis dahin anekdotisch behandelten MWE ins Zentrum rückt.

In Abbildung 3 ist eine Baumdarstellung der MWE-Struktur nach Maurice Gross bzw. der Gross-Schule dargestellt.

MWEs heißen bei Gross „expressions figées“. Sie unterteilen sich in satzförmige „phrases figées“ und nicht-satzförmige „locutions figées“[12]. Die satzförmigen untergliedern sich in Sprichwörter, Routineformeln sowie Klassen nach Stützverb (A = *avoir*, E = *être*, C = andere Verben). Die nicht-satzförmigen fächern sich nach Wortarten auf, also in „locutions nominales“, „locutions verbales“ etc. sowie komplexe Determinierer.

Die Äste in Abbildung 3 sind in den Arbeiten von Gross und dem LADL detailliert ausgearbeitet worden. Für die nicht-satzförmigen siehe vor allem die Monographie von Gaston Gross „Les expressions figées en français“ ([G. Gross 1996]) sowie das Kapitel zu Determinierern in [G. Gross 2012]. Die satzförmigen *Phrases figées* hat Maurice Gross selbst detailliert bearbeitet, siehe insbesondere „Les phrases figées en français“ [M. Gross 1993a] sowie [M. Gross 1982], [M. Gross 1988]. Diese sehr aufwändigen, manuell erstellten Bestandsaufnahmen wurden von seinen Schülerinnen und Schülern fortgeführt und bildeten den Ausgangspunkt der bis heute aktiven Gross-Schule. Von besonderem Interesse für diese Arbeit ist die Klassifikation von Michel Mathieu-Colas ([M. Mathieu-Colas 1996]), der über 700 Typen nominaler MWEs im Französischen identifiziert und beschreibt.

[12] Soweit ersichtlich, verwendet Gross keine eigene Dachbezeichnung für den nicht-satzförmigen Bereich. Ich verwende „locutions figées“ hier als Sammelbegriff zur Darstellung.

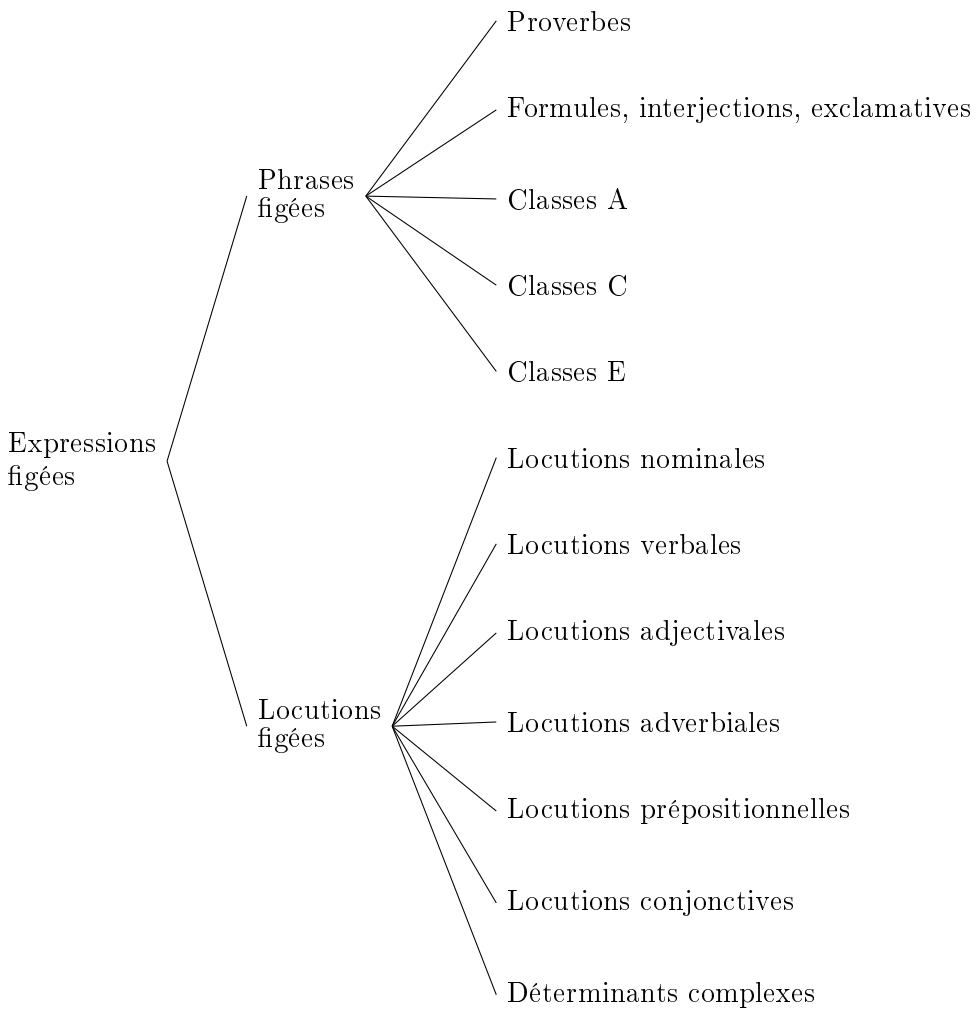

Abbildung 3: Klassifikation nach Gross bzw. der Gross-Schule

Im Laufe der Jahre sind aus der Gross-Schule weitere Taxonomien für MWEs hervorgegangen, von denen ich wichtige Vertreter im Folgenden vorstelle. Direkt auf Gross aufbauend legen Guenthner & Blanco eine viergliedrige Typologie vor, die die lexikon-getriebene Sicht der Gross-Schule zu einer handhabbaren Obergliederung bündelt und erkennbar ein Update der Gross'schen Taxonomie darstellt, mit dem Fokus auf Kompaktheit und Implementierbarkeit:

Franz Guenthner und Xavier Blanco (2004): Die Klassifizierung von [F. Guenthner 2004] unterscheidet vier Typen von MWEs, wobei syntaktische und lexikologische Kriterien im Vordergrund stehen (Wortarten und deren Beweglichkeit im Satz). Die ersten beiden Kategorien umfassen die große Mehrzahl der MWEs, nämlich solche, die als eine (oder mehrere) der klassischen Wortarten Noun, Verb,

Adjective, Adverb oder Preposition in einem Lexikon einen Platz finden können. Diese werden weiter nach dem Kriterium unterschieden, ob sie keine interne Struktur aufweisen und als ein einzelnes lexikalisches Objekt behandelt werden müssen („Compounds“), oder ob sie über eine interne Struktur und einen echten semantischen Kopf verfügen („Collocations“) und damit als zwei oder mehrere lexikalische Objekte behandelt werden müssen. Die anderen beiden Kategorien lassen sich im Gegensatz zu den erstgenannten nicht in Wortarten einteilen, sondern sind auf höherer Ebene („Frozen Texts“) oder niedrigerer Ebene („Grammatical Multi-Lexemic Expressions“) als zusammengehörig zu betrachten. Die Struktur ist in Abbildung 4 dargestellt. Im Detail ist die Klassifikation folgendermaßen aufgeschlüsselt:

Compound Expressions sind zusammengesetzte Wörter unterschiedlicher Wortarten, die als zusammengehörige Einheit in ein Lexikon aufgenommen werden müssen. Sie entsprechen damit den bisher mit MWE bezeichneten Einheiten. Im Falle einer Übersetzung müssen diese Begriffe als Ganzes und ohne Rücksicht auf ihre Bestandteile oder innere Struktur übersetzt werden. Je nach Sprache kann diese Zusammensetzung entweder durch einen Bindestrich, einen *empty separator* oder ein Leerzeichen realisiert werden. Compound Expressions lassen sich weiter untergliedern in Compound Nouns (CNs), Compound Verbs, Compound Adverbs, Compound Adjectives und Compound Determiners. Diese Unterscheidung lehnt sich in ihrer Untergliederung an die bereits von Maurice Gross vorgeschlagenen Kategorien an.

Collocations müssen als zwei (oder mehr) Objekte behandelt werden, deren innere Struktur etwa bei der Übersetzung aufrechterhalten bzw. berücksichtigt werden muss. Eine weitere Unterteilung erfolgt in Frozen Modifiers, d. h. Nomen, Verben und Adjektive, die systematisch durch ein mehr oder weniger fixes Vokabular semantisch modifiziert werden können („starker Raucher“) und Support Verb Structures, d. h. Strukturen, in denen das entsprechende Nomen ausschließlich mit einem bestimmten Hilfsverb kombiniert werden kann („einen Vortrag halten“).

Frozen Texts (d. h. idiomatisierte feste Ausdrücke wie Sprichwörter) bilden auf einer Ebene höher als der Wortebene Einheiten, d. h. es handelt sich um Teil- oder ganze Sätze wie Sprichwörter oder stehende Wendungen.

Grammatical multi-lexemic expressions beinhalten weitere, allerdings schwerer zu fassende MWEs. Sie lassen sich in zwei Untergruppen aufteilen:

Empirical Grammatical Expressions Statistisch signifikante Kookurrenzen, die sich keiner der bisher genannten Klassen zuordnen lassen, etwa temporale Konstruktionen wie „has been“ oder Konjunktionen wie „either... or“)

Theoretical Grammatical Expressions Zusammen auftretende Ausdrücke wie „red t-shirt“, bei denen ein oder mehrere der Konstituenten zwar eine größere Varianz haben, die aber semantisch fixiert [red, blue, green... COLOR] [t-shirt, sweater, pants... CLOTHING] ist. Diese Formen müssen daher als Regeln in einem Lexikon kodiert werden.

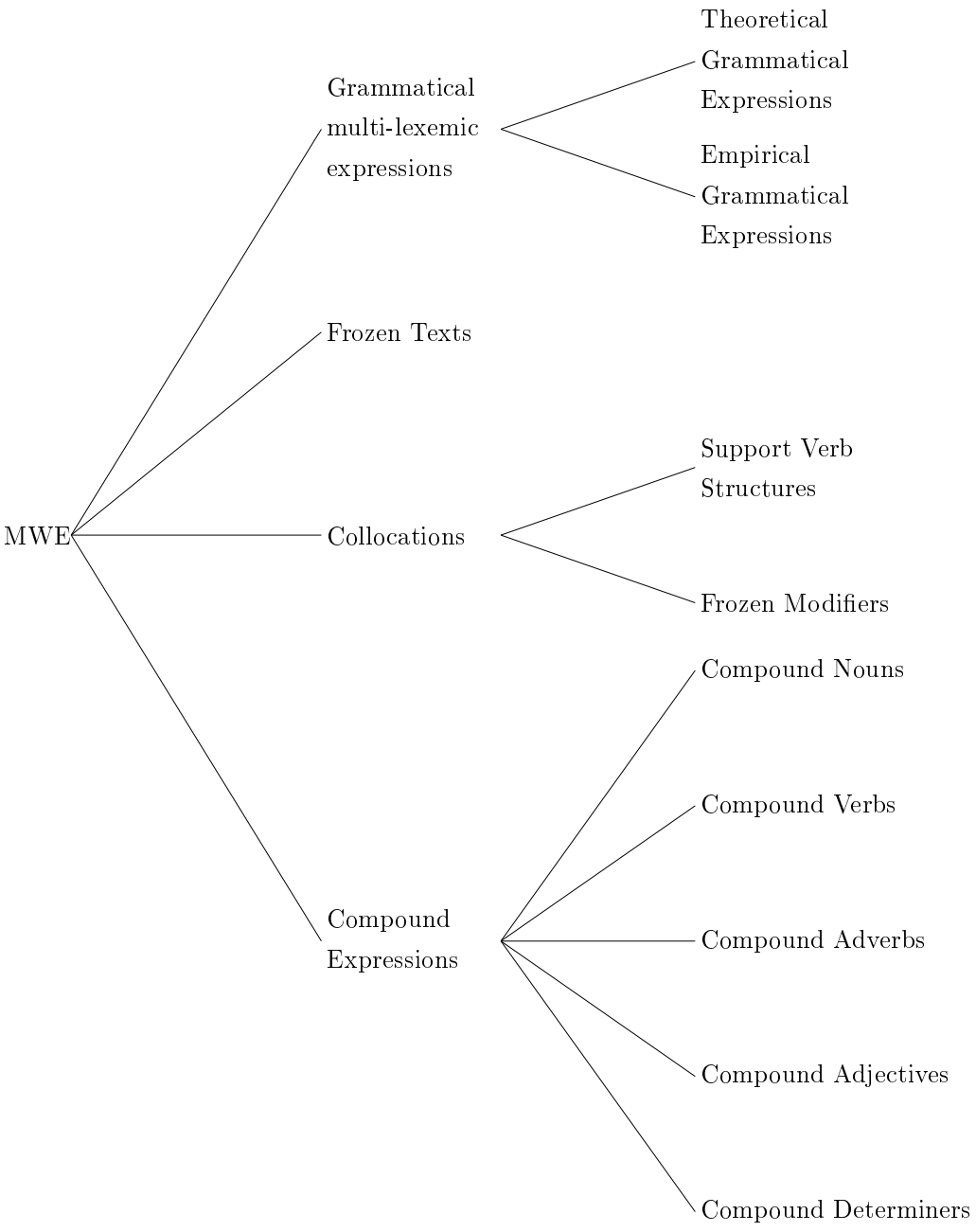

Abbildung 4: Klassifikation nach Guenthner und Blanco (2004)

Éric Laporte (2018): Als jüngere Aktualisierung der auf Gross basierenden Klassifikationen schlägt Éric Laporte in [É. Laporte 2018] eine merkmalsgeleitete Sicht vor: Statt vorab gewählte, teils unscharfe Kategorien zu reproduzieren, plädiert er für klar entscheidbare, computergerechte Merkmale und eine Gliederung, die sprachübergreifend tragfähig ist:

> Some features are fuzzy and imprecise, that is, it is difficult to tell which MWEs have them. In resulting classifications, assignment of MWEs to classes is less reliable than it could be, and this is detrimental to computational use. Other features are more clear-cut and potentially more useful, but have not made their way to computational-linguistic literature yet. Another requirement for a convenient classification is that its outline be suitable for many languages.
>
> [É. Laporte 2018, 143f]

Auf dieser Basis präsentiert er zwei kompakte Alternativen:

- Eine Sag-nahe Taxonomie mit der obersten Unterscheidung *lexikalisiert* vs. *nicht lexikalisiert*. Die lexikalisierten Einheiten teilen sich weiter in MWEs *ohne Support-Verb* (z. B. mehrwortige Nomen, Adverbien und Adjektive sowie verbale Idiome) und in *Funktionsverbgefüge mit Support-Verb* (Support Verb Construction (SVC) bzw. Light Verb Construction (LVC)).
- Eine erweiterte Variante, in der *Kopulaverben* als *Support-Verben* mitgezählt werden. Damit fallen auch Prädikationen mit Adjektiv bzw. Präpositionalphrase als Kern (*be angry*, *be on time*) unter SVC. Diese Variante ist in Abbildung 5 dargestellt.

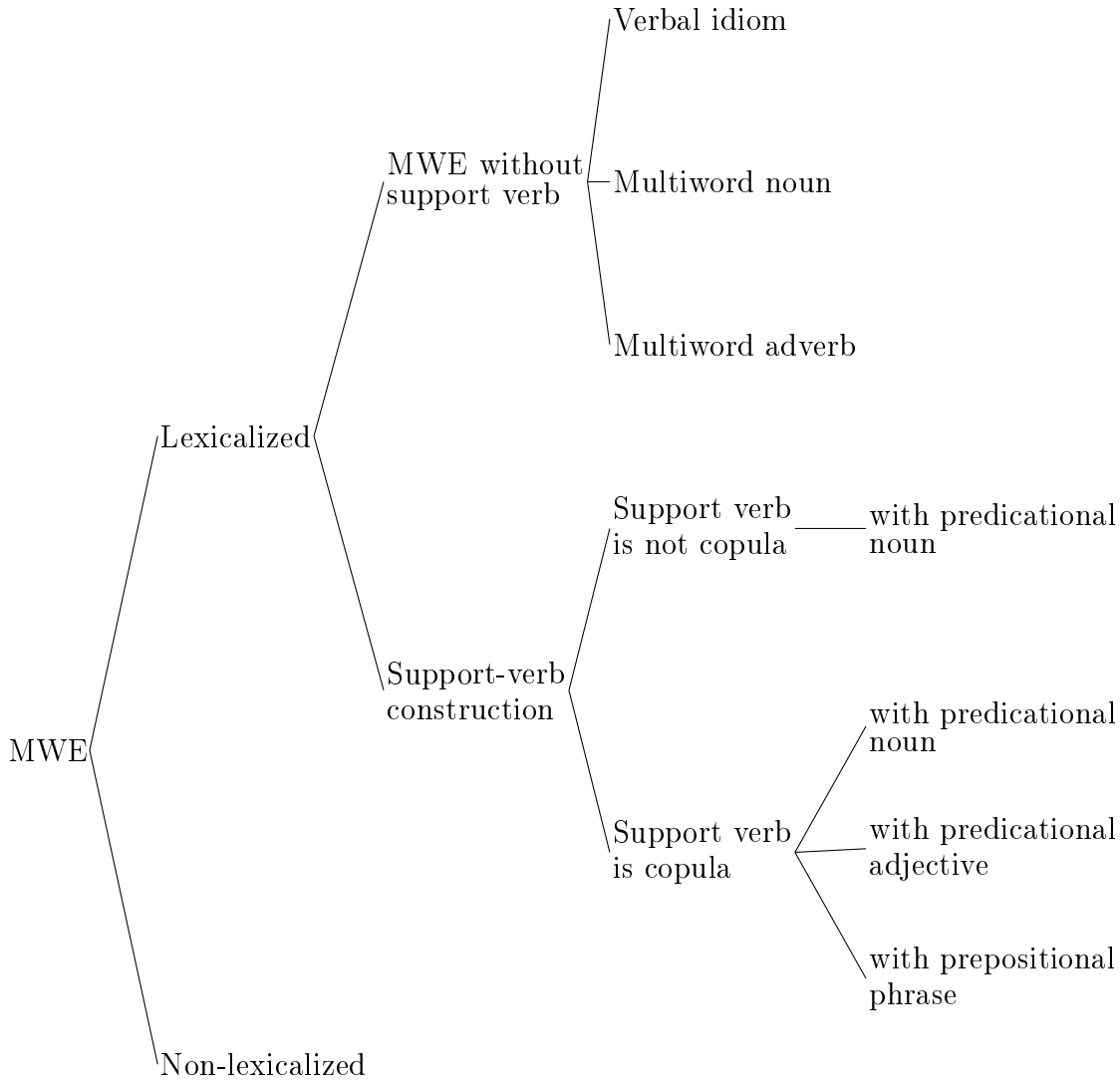

Abbildung 5: Klassifikation nach Laporte (2018)

2.3.3 Vergleich der Klassifizierungen

Die beschriebenen Klassifikationen decken ein breites Spektrum ab. An erster Stelle fällt auf, dass die Taxonomien sich entlang unterschiedlicher Leitfragen bewegen: Sag et al. ordnen MWE entlang des Grades der Fixierung, Gross anhand von Wortarten nach vorheriger Unterscheidung satzförmig vs. nicht-satzförmig. Laporte unterscheidet primär lexikalisierte und nicht lexikalisierte Formen. Ebenfalls unterscheiden sich die Ansätze hinsichtlich ihrer Formalisierung und damit möglichen Implementierung. Kunze und Lemnitzers Klassifikation ist noch nicht abgeschlossen, aber klar lexikographisch orientiert. Die Klassifikation von Sag et al. ist einflussreich, aber zu grob

für eine Implementierung, die Klassen sind über graduelle Eigenschaften beschrieben, ohne klare Entscheidungskriterien. Formal ausdefiniert sind insbesondere die Grossschen Klassen sowie darauf aufbauend Guenthner und Blanco (inklusive regelbasierter „theoretischer grammatischer Ausdrücke“) beziehungsweise mit binärem Entscheidungsbaum und unter Einbeziehung von Kopulaverben die Klassifikation von Laporte. Weitere augenfällige Unterschiede betreffen die Einordnung von Kollokationen und SVC, die ich an dieser Stelle jedoch nicht vertiefen werde, da sie nicht dem Fokus der Arbeit entsprechen.

Auf dieser Basis behandle ich als Nächstes die Unterklasse der nominalen MWEs und ihre konkrete Ausprägung in den Taxonomien.

2.4 Die Unterklasse nominaler MWEs

Obwohl alle Klassifikationen nominale MWEs berücksichtigen, unterscheiden sie sich in der konkreten Abbildung. Bei Sag et al. erscheinen sie namentlich nur als „(semi-fixed) Compound Nouns“. Bei Kunze und Lemnitzer verteilen sie sich auf „Phraseme“, „Kollokationen“ und „mehrgliedrige Komposita“. Gross führt sie explizit als „locutions nominales“. Guenthner und Blanco listen sie als „Compound Nouns“, Laporte als „Multiword Nouns“.

Nominale MWEs sind in allen Klassifikationen erfasst und bilden keinen Grenzfall. Sie stellen die zahlenmäßig stärkste und produktivste Klasse komplexer lexikalischer Einheiten dar und nehmen entsprechend erheblichen Raum in Texten ein (vgl. Abschnitt 3.1.3). Als Teile von Nominalphrasen tragen sie wesentlich zur Textbedeutung bei und sind für zahlreiche computerlinguistische Anwendungen relevant (vgl. Abschnitt 2.6).

Allerdings liegt der Schwerpunkt der internationalen MWE-Forschung seit einiger Zeit tendenziell auf verbalen Einheiten, etwa in der *Shared Task* von PARSEME zu *Verbal Multiword Expressions* [C. Ramisch 2020]. Nominale MWEs sind daher vergleichsweise untererforscht und in vielen Ressourcen unterrepräsentiert (vgl. Abschnitt 3.1.2). Ihr scheinbar einfacherer Charakter täuscht, denn idiomatische Effekte wie semantische Intransparenz und der fließende Übergang zur Lexikalisierung stellen auch hier erhebliche Herausforderungen dar.

Definition nominaler MWE Nominale MWEs sind konventionalisierte, mehrwortige Einheiten, die als Nominalphrasen oder als nominale Teile davon fungieren. Sie sind in der Regel kontinuierlich realisiert. Je nach MWE sind interne Modifikatio-

nen wie Flexion oder adjektivische Einfügungen bedingt möglich. Als Listeme sind sie im Lexikon zu erfassen, wenn mindestens eines der folgenden Merkmale vorliegt.

1) **syntaktisch oder operationell idiomatisch** Mindestens ein Bestandteil oder eine interne Relation ist festgelegt. Ersetzung, Umstellung oder Tilgung führt zu unidiomatischen oder ungebräuchlichen Formen.
   Beispiel: *agujero negro*, nicht **agujero muy negro*.

2) **semantisch idiomatisch** Die Gesamtbedeutung ist nicht vollständig kompositionell erschließbar.
   Beispiel: *príncipe azul* für Traummann.

3) **distributionell oder pragmatisch idiomatisch** Die Kombination ist gegenüber regulären Alternativen eindeutig präferiert oder exklusiv und zeigt ein eigenes distributionsbezogenes oder selektionsbezogenes Profil.
   Beispiel: *piedra angular* nicht frei austauschbar mit **roca angular*.

Obwohl die Definition vergleichsweise formal gefasst ist, bleiben die Grenzen in der Praxis häufig fließend, besonders mit Blick auf den Grad der Konventionalisierung. Ich nutze die Definition daher nicht als harte Filterbedingung, sondern als Heuristik und als Referenzrahmen. Für LEMoN verfolge ich einen recall-orientierten Ansatz. Nominale Kandidaten werden breit erfasst, auch reguläre transparente und nicht lexikalisierte Nominalphrasen. Potenzielle Übererfassung kann in nachgelagerten Schritten gewichtet und gefiltert werden. Wichtiger ist, dass keine relevanten Einheiten verloren gehen.

Nach der Definition des Arbeitsgegenstandes nominaler MWEs folgt zunächst ein Überblick über die historische Entwicklung der MWE-Forschung. Daran anschließend werden aktuelle Forschungsstränge sowie die besondere Relevanz des Themas für die Verarbeitung dargestellt. In Kapitel 3 wird die Definition auf das Spanische angewendet und mit Blick auf dessen sprachspezifische Eigenheiten ausgeführt.

2.5 Geschichte der MWE-Forschung

Mit dieser wissenschaftsgeschichtlichen Darstellung der MWE-Forschung werden drei Ziele verfolgt:

1) Darstellung des historischen Hintergrunds der MWE-Forschung mit besonderem Schwerpunkt auf die romanischen Sprachen und insbesondere das Spanische.

2) Nachzeichnung methodischer Verfahren und ihrer Entwicklung über die Zeit.

3) Verdeutlichung der besonderen Relevanz von MWEs für die computerlinguistische Sprachverarbeitung - auch im Hinblick auf moderne Trends wie Large Language Models (LLMs).

Der Überblick ist nicht als vollständige internationale Geschichte der MWEs-Forschung angelegt, sondern konzentriert sich auf die für diese Arbeit relevanten Entwicklungen, insbesondere im Spanischen.

2.5.1 Von der Antike bis 1960

Die Beschäftigung mit Sprichwörtern, also die Parömiologie, gilt gemeinhin als jene sprachwissenschaftliche Disziplin, aus der sich die moderne Phraseologie und in der Folge auch die MWEs-Forschung entwickelt haben. Sie ist zweifellos ein zentraler Ausgangspunkt, insbesondere für die empirische Beschäftigung mit festen Wortverbindungen. Daneben existieren jedoch weitere Traditionslinien, etwa in der Terminologieforschung sowie in pragmatischen, lexikographischen, syntaktischen und semantischen Arbeiten, die sich ebenfalls mit mehrgliedrigen Einheiten befassten.

Frühe Reflexionen über feste sprachliche Wendungen, etwa Sprichwörter, Redensarten und feste Formeln, finden sich bereits in der antiken Rhetorik und Grammatik, doch systematische Ansätze entstehen erst nach der institutionellen Begründung der Sprachwissenschaft in der Neuzeit. Und obwohl bereits das von der Real Academia Española zwischen 1726 und 1739 herausgegebene *Diccionario de autoridades* ([Real Academia Española 1726 1739]) schon auf dem Titelblatt ankündigt, auch *proverbios* und *refranes* zu beinhalten, dauert es noch bis 1969, ehe Julio Casares mit seiner Definition von MWEs als „locuciones“ erstmals explizit die Kriterien Polylexikalität, Fixiertheit und Idiomatizität benennt. Als erste umfassende Arbeit zu spanischen MWEs, die „eine sachliche und methodisch kohärente metatheoretische Reflexion der Beschreibung des spanischen phraseologischen Bestandes [...] zugrunde legt“ ([M. T. Zurdo 2007, 704 f]), gilt jedoch die Studie von Alberto Zuluaga [A. Zuluaga 1980].

2.5.2 Distributionalismus und die Gross-Schule (1960-1990)

Ab den 1960er Jahren tritt neben die philologisch-lexikographisch geprägte Tradition eine formal-empirische Linie: die Pariser Gross-Schule - die für diese Arbeit von

besonderem Interesse ist. Ihr Ursprung liegt jedoch nicht in Frankreich, sondern in den USA, im Umfeld der amerikanischen distributionalistischen Tradition.[13]

Der französische Linguist Maurice Gross (1934-2001) arbeitete in den 1960er-Jahren zeitweise bei Zellig S. Harris (1909-1992) an der University of Pennsylvania. Harris, der an die strukturalistische Linie Leonard Bloomfields (1887-1949) anschloss (vgl. [L. Bloomfield 1933]), verstand Grammatik nicht als abstraktes Regelwerk, sondern als ein System, das sich aus den beobachtbaren Verteilungen sprachlicher Einheiten in ihren Kontexten erschließen lässt. Für die Analyse dieser Distributionen entwickelte er frühe formale Verfahren (vgl. [Z. S. Harris 1963], [Z. S. Harris 1968], [Z. S. Harris 1991]). Damit legte er die Grundlage für den amerikanischen Distributionalismus, in dem er heute als zentrale Figur gilt.

> The distribution of an element is the total of all environments in which it occurs, i.e. the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements.
>
> [Z. S. Harris 1963, 15f]

Harris war zugleich ein prägender akademischer Lehrer. Aus seiner Denkschule gingen jedoch sehr unterschiedliche Strömungen hervor. Sein Doktorand Noam Chomsky führte die distributionellen Ideen in eine ganz andere Richtung: zur *Generativen Grammatik*, die Sprachstruktur aus universalen Regeln ableitet und dabei auf einer strikten Trennung von Syntax und Semantik insistiert (vgl. [N. Chomsky 1957], [N. Chomsky 1965]). Damit verlagerte sich der Fokus von empirisch beobachteten Distributionen hin zu theoretischen Universalien.

Maurice Gross hingegen schlug den entgegengesetzten Weg ein. Er knüpfte stärker an Harris' empirische Tradition an und übertrug deren Konzepte auf lexikalische Inventare. Im Nachruf „In Memoriam Maurice Gross" formuliert es Éric Laporte so:

> There is a smooth continuum from the thought and work of Zellig Harris to that of Maurice Gross, and it is sometimes difficult to differentiate between their respective contributions.
>
> [É. Laporte 2005a, 1]

Auf dieser Basis entwickelte er die Theorie der (untrennbaren) *Lexikon-Grammatik*, die die enge Verschränkung von Lexikon und Grammatik systematisch dokumentiert.

[13] Zur Verortung der strukturalistischen Tradition im europäischen Kontext vgl. [J. Albrecht 2000] sowie für die spanische Grammatikographie [C. Timm 2010].

Im Zuge der *Linguistics Wars*[14] etablierte sich die generative Linie als dominanter Mainstream. Parallel dazu entstand jedoch am Pariser Laboratoire d'Automatique Documentaire et Linguistique (LADL) unter der Leitung von Gross eine Schule, die den distributionellen Blick von Harris mit lexikalischer Präzision verband und damit eine dauerhafte Alternative zur generativen Grammatik ausbildete.

Anders als die übliche Trennung von Lexikon und Grammatik betrachtet die Lexikon-Grammatik beides als eng verflochtenes System, dessen Regeln empirisch zu dokumentieren sind. Das Kernelement der Gross-Schule bilden „elementare Sätze", also prototypische Satzrahmen mit argumentativen Leerstellen. Ihre Kombinatorik wird in Lexikon-Grammatik-Tabellen (LG-Tabellen) erfasst, in die einfache und rekursiv-komplexe lexikalische Einheiten eingesetzt werden können. Letzteres führt direkt zum Thema dieser Arbeit: die explizite Behandlung von MWEs über alle Fixierungsgrade hinweg - einerseits als Listeme in LG-Tabellen, andererseits mittels lokaler Grammatiken für regulär produktive, aber nicht völlig freie Bildungen.

> Avec les expressions figées, la situation est différente; les formes que nous étudierons ont toujours été considérées comme des exceptions. Aucune règle n'a donc été envisagée pour elles. Les exemples ont été confinés dans des glossaires spécialisés où des anecdotes leur sont attachées. Nous montrerons que nous avons affaire à un phénomène d'envergure, masqué par une absence d'études, due à des *a priori* sur la nature du langage.[15]
>
> [M. Gross 1982, 151]

Somit konstatiert Gross: MWEs sind keine Ausnahmen, die anekdotisch aufgeführt werden können, sondern sie sind ein wesentlicher Bestandteil des Sprachinventars und sollten Gegenstand einer systematischen Beschreibung sein.

Obwohl die Arbeiten des LADL zunächst international weniger sichtbar blieben (u.a. sprach- und formatbedingt), entfalteten sie eine nachhaltige Wirkung und begründeten eine eigenständige Schule, die bis heute fortwirkt. Und die Gross-Schule

[14] Als „Linguistics Wars" werden die öffentlichen Auseinandersetzungen innerhalb der Generativen Grammatik in den 1960er- und 1970er-Jahren bezeichnet (vgl. [R. A. Harris 1993]).

[15] Bei den festen Ausdrücken verhält es sich anders: die Formen, die wir untersuchen, sind immer als Ausnahmen angesehen worden. Daher wurden für sie auch keine Regeln erstellt. Die Beispiele wurden in spezialisierte Glossare abgeschoben und dort anekdotisch illustriert. Wir werden (jedoch) zeigen, dass wir es mit einem weiter verbreiteten Phänomen zu tun haben, das durch die Abwesenheit von Studien verdeckt wird, die (wiederum) gewissen Vorannahmen über die Natur der Sprache (langage!) geschuldet sind. (Übersetzung des Verfassers)

ist auch nicht die einzige Tradition, die sich sich systematisch mit MWEs auseinandergesetzt hat, geblieben: Parallel argumentierte zur gleichen Zeit auch der russisch-kanadische Linguist Igor Mel'čuk, dessen *Erklärend-Kombinatorisches Wörterbuch* viele Anknüpfungspunkte an die Gross-Schule und die Behandlung von MWEs in Lexika bietet:

> Se dice muy a menudo que el ser humano habla con palabras; ello presupone que, para hablar bien una lengua, basta con dominar el léxico (=las palabras) y la gramática (=la sintaxis + la morfología). Pero es falso: el lexico y la gramatica son necesarios pero distan mucho de ser suficientes.
>
> [I. Mel'čuk 2006, 3][16]

2.5.3 Von RELEX bis PARSEME (1990–2020)

Seit etwa 1990 ist eine starke Zunahme an phraseologischer Forschung sowohl in der klassischen Linguistik als auch in der Computerlinguistik zu beobachten. Dies hängt einerseits mit der Institutionalisierung der Phraseologie als eigenständigem Zweig der Sprachwissenschaft zusammen. Andererseits führte die „Wiederentdeckung" statistischer Methoden in der Computerlinguistik durch [P. F. Brown 1990] sowie die gestiegene Verfügbarkeit von Computern und elektronischen Textressourcen dazu, dass die Unzulänglichkeit von einfachen Wörtern als alleinigen syntaktischen Atomen stärker in den Fokus rückte.

Zur Förderung der internationalen Forschung an der Lexikon-Grammatik gründete Maurice Gross 1991 das lose RELEX-Network[17]. Ziel dieser Vereinigung war die umfassende lexikalische Erschließung der jeweiligen Mitgliedssprachen nach einem einheitlichen Schema. Mithilfe der hierbei gewonnenen Daten sollten die von Gross skizzierten weiteren Arbeitsschritte vertieft werden. Die Beschäftigung mit den Theorien der Gross-Schule beginnt in Deutschland und Spanien jeweils etwa Mitte der 1990er Jahre. Im deutschsprachigen Raum wurde dieser Ansatz vor allem durch Franz Guenthner am Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig-Maximilians-Universität München (LMU) vermittelt, ein

[16] „Es wird oft behauptet, dass der Mensch mit Wörtern spricht; das impliziert, dass es, um eine Sprache gut zu sprechen, ausreichen würde, das Lexikon (=die Wörter) und die Grammatik (=die Syntax und die Morphologie) zu beherrschen. Aber das ist falsch: Lexikon und Grammatik sind notwendig, aber weit davon entfernt, ausreichend zu sein." (Übersetzung des Verfassers)

[17] `http://infolingu.univ-mlv.fr/english/Relex/Relex.html`, Stand: 10.10.2025

Beispiel ist das Projekt CISLEX (vgl. [F. Guenthner 1994, S. Langer 1996]). In Spanien finden sich RELEX-Partner sowohl an der Universitat Autònoma de Barcelona (UAB) mit Xavier Blanco Escoda und den spanischen DELA-Wörterbüchern (vgl. [X. Blanco Escoda 2001]) als auch an der Universidad A Coruña mit Margarita Alonso Ramos und dem Diccionario de Colocaciones Españoles (DiCE) (vgl. [M. Alonso Ramos 2010]).

Zwar spielte das RELEX-Netzwerk zu Lebzeiten von Maurice Gross eine bedeutende Rolle in der Entwicklung der internationalen Lexikon-Grammatik, ist heute aber - scheinbar - weitgehend inaktiv. Zwar sind noch einzelne Mitglieder aktiv in der MWE-Forschung, aber gemeinsame Aktivitäten sind nicht ersichtlich.

Einen guten Überblick über die allgemeine Entwicklung und den Stand der spanischen Phraseologieforschung geben [M. T. Zurdo 2007] und [M. García-Page Sánchez 2008]. Mit den folgenden Worten fasst María Teresa Zurdo den Stand der Forschung in der spanischsprachigen Linguistik zum Thema MWE zum Ende dieser Phase zusammen:

> In der Hauptsache handelt es sich um wissenschaftliche Aufsätze, die sich mit einzelnen Eigenschaften des spanischen phraseologischen Systems befassen. Letzten Endes dominiert die je nach Forschungsziel einzelsprachlich und/oder kontrastiv orientierte Behandlung von morphologischen, syntaktischen, lexikalisch-semantischen, pragmatischen, textuellen oder kulturellen Merkmalen fester Wortverbindungen.
>
> [M. T. Zurdo 2007, 704]

2013 wurde das internationale, interdisziplinäre Netzwerk *PARsing and Multiword Expressions* (PARSEME)[18] gegründet. Als mittlerweile fest etablierte Instanz veranstaltet es regelmäßig MWE-bezogene Konferenzen, deren Proceedings in dieser Arbeit häufig zitiert werden. Die Mitglieder von PARSEME haben sich zudem auf ein lesenswertes Memorandum of Understanding geeinigt, vgl. [COST 2012]. Darüber hinaus nutzt das Netzwerk die Community, um mit Surveys[19] die aktuelle Datenlage sichtbarer zu machen oder mit Shared Tasks größere Aufgaben europaweit gemeinsam zu bearbeiten (vgl. [A. Savary 2017]).

[18] `https://typo.uni-konstanz.de/parseme/`, zuletzt abgerufen am: 10.10.2025

[19] vgl. `https://typo.uni-konstanz.de/parseme/index.php/2-general/159-survey`, zuletzt abgerufen am: 10.10.2025

2.5.4 LLM-Zeitalter (seit 2020)

Seit Beginn der 2020er Jahre üben die mittlerweile allgegenwärtigen LLMs einen ambivalenten Effekt auf die Forschung zu MWEs aus. Auf der einen Seite haben sie es geschafft, bei ausreichendem Trainingsmaterial auch die berüchtigt schweren MWEs über Sprachgrenzen hinweg überzeugend zu nutzen und zu übertragen, eine Leistung, die noch vor kurzer Zeit als Science Fiction galt. Und dennoch haben sie erhebliche Defizite bei der Identifikation, Klassifikation und dem linguistischen Verständnis dieser Einheiten.

Eine Bestandsaufnahme bieten die Vorträge von Vered Shwartz mit den Titeln „Reality Check: Natural Language Processing in the era of Large Language Models“ von der *Munich NLP 2023* sowie „A long hard look at MWEs in the age of Language Models“[20] . Unter diesen auf den ersten Blick widersprüchlichen Vorzeichen beginnen also auch im Jahr 2025 noch viele Arbeiten, die sich mit der Verarbeitung von MWEs auseinandersetzen, mit einem wie folgend typischen Einstieg:

> Multiword expressions (MWE) are idiosyncratic word combinations that constitute both major challenges and interests in Natural Language Processing.
>
> [L. Castro 2025, 32]

Es gilt die Problemformel des mittlerweile mehr als 20 Jahre alten zentralen Artikels für die MWE-Forschung von [I. A. Sag 2002]:

> We believe the problem of MWEs is critical for NLP, but there is a need for better understanding of the diverse kinds of MWE and the techniques now readily available to deal with them.
>
> [I. A. Sag 2002, 3]

Entsprechend schreibt auch Carlos Ramisch in seiner 2023 erschienenen Monographie „Multiword expressions in computational linguistics“:

> Fast progress in language technologies is driven by the research community in natural language processing (NLP), of which I am part. My personal holy grail in NLP would be achieving robust and precise representation

[20] `https://www.youtube.com/watch?v=mTQM49YkjXU`; `https://www.youtube.com/watch?v=0pc46DDVsP4`. Zuletzt abgerufen am 10.10.2025.

> and processing of multiword expressions [...] in computational linguistic models and applications. Much water has flown under the bridge between the famous pain-in-the-neck paper (Sag et al. 2002) and the latest edition of the multilingual PARSEME shared task (Ramisch et al. 2020). For the last 15 years, I have been eating, sleeping and breathing multiword expressions, witnessing and contributing to significant advances in how we deal with them in text processing. [...] Because the phenomenon is so rich, I have mostly succeeded in coming up with new ways to illustrate what multiword expressions are and why it is important to work on this subject [...]
>
> [C. Ramisch 2023, 2]

Neben der (computer-)linguistischen Behandlung von MWEs finden sich auch Arbeiten aus angrenzenden Disziplinen. So untersuchen neurolinguistische und psycholinguistische Studien die kognitive Verarbeitung fester Wortverbindungen (vgl. [A. Siyanova-Chanturia 2017], [S. Jiang 2020]). Darüber hinaus gibt es eine Vielzahl domänenspezifischer Ansätze, die von der Verständlichkeit biomedizinischer Texte [S. Bagdasarov 2024] über Diskursanalysen zu neuen Wendungen in der COVID-19-Pandemie [A. Ladilova 2024] bis hin zu Spezialterminologien wie der spanischen Stierkampfsprache reichen [E. Gutiérrez Rubio 2024].

2.6 Relevanz

Die Bedeutung von MWEs ist im Verlauf dieser Arbeit bereits mehrfach angesprochen worden. An dieser Stelle sollen die zentralen Argumente gebündelt zusammengefasst werden.

Zum einen stellen MWEs einen erheblichen Teil des lexikalischen Inventars dar und werden seit Jahrzehnten als „Bottleneck“ der Sprachverarbeitung bezeichnet (vgl. [COST 2012]). Besonders nominale MWEs gelten aufgrund ihrer Produktivität, insbesondere in den sich kontinuierlich ausdifferenzierenden und ausspezifizierenden Fachsprachen, als „the bulk of the lexicon of languages“ ([M. Gross 1986, 4]).

Zum anderen kommt den MWEs auch eine inhaltliche Schlüsselfunktion zu: Sie sind nicht nur semantische Einheiten, sondern tragen häufig die zentralen Konzepte eines Textes. [W. Zhang 2008] zeigen, dass kurze Mehrwortausdrücke oft die Hauptthemen markieren, während längere Unterthemen repräsentieren.

Schließlich sind sie für praktisch alle Anwendungen der Sprachverarbeitung unverzichtbar. Sie müssen als Einheiten erkannt, analysiert und generiert werden, andernfalls entstehen systematische Fehler - etwa in der maschinellen Übersetzung, im Information Retrieval oder in der Textklassifikation. Wie Sag et al. in [I. A. Sag 2002] betonen, eröffnet erst die präzise Identifikation von MWEs auch eine zuverlässigere Erkennung ihrer Bestandteile und Kontexte. Und das PARSEME-Memorandum hebt hervor, dass NLP-Systeme zugleich sprachliche Präzision, Sprachspezifika und Effizienz im Umgang mit großen Datenmengen gewährleisten müssen:

> One of the key problems to be overcome in order to meet all of these requirements simultaneously are multi-word expressions [...]
>
> [COST 2012, 3f]

Damit ist auch die Relevanz doppelter Natur: Theoretisch sind MWEs ein Baustein für empirisch belastbare Grammatik- und Lexikonmodelle, praktisch sind sie Ressourcen für präzisere und robustere NLP-Anwendungen. Vor diesem Hintergrund richtet sich das folgende Kapitel auf die Charakteristika nominaler MWEs im Spanischen und ihre Einordnung in die hispanistische Forschungstradition.

3 Nominale MWE im Spanischen

Intelijencia, dame el nombre esacto de las cosas!

Juan Ramón Jiménez

Dieses Kapitel behandelt die charakteristischen Eigenschaften nominaler MWEs im Spanischen. Nach einem an den Einstieg des vorigen Kapitels angelehnten Überblick über allgemeine Segmentierungsphänomene im Spanischen folgen Typologien nominaler Komposita und MWEs sowie deren Kompositionsverfahren. Abschließend richtet sich der Blick auf morphologische Prozesse und ihren Einfluss auf die jeweilige Oberflächenform.

Auch im Spanischen treten verschiedene Segmentierungsprobleme auf, die den intuitiven Wortbegriff früh herausfordern. Zum einen wird ein Verb in der Regel von seinen Argumenten getrennt geschrieben, im Imperativ jedoch werden Verb und Argumente zu einem Wort ohne Separatoren zusammengezogen:

1) *Me lo da*

2) *Dámelo*

Neben diesen verbalen Phänomenen gibt es auch im Bereich der Nominalbildung sprachspezifische Eigenheiten zu beachten. So gibt es eine Klasse, die den deutschen agglutinierten Nominalkomposita ähnlich ist, jedoch sehr viel weniger produktiv oder bedeutsam:

1) *sacacorchos*

2) *medianoche*

Obwohl diese auch unter eine weite Definition von MWE fallen könnten, sind die agglutinierten Bildungen eher ein Randphänomen. Für LEMoN relevant sind hingegen die syntagmatischen, mehrgliedrigen Einheiten, die orthographisch getrennt erscheinen, wie in den Abschnitten 2.1.3 für MWE und 2.4 für nominale MWE definiert:

1) *diente de león*

2) *hombre lobo*

3) *ley de propiedad intelectual*

4) *amor platónico*

5) *alta mar*

3.1 Zielbereich und sprachlicher Rahmen

Für diese Arbeit sind nominale MWEs im weitesten Sinne relevant. Zwar fokussiert die Untersuchung auf MWE im Sinne der Definition aus Abschnitt 2.1.3, der hier verfolgte *recall-orientierte* Ansatz unterscheidet jedoch nicht nach dem Grad der Idiomatizität. Er identifiziert neben stark oder schwach fixierten MWEs auch Kollokationen und reguläre Bildungen. Dies ist bewusst so gewählt: Eine Aufnahme ins Wörterbuch zu viel - selbst wenn der Lexikalisierungsstatus diskutabel ist - ist methodisch weniger problematisch als das Fehlen einer idiomatischen MWE.

LEMoN ist als Lexikon spanischer nominaler MWEs konzipiert. Terminologisch werden diese auf Deutsch als „Nominalkomposita“, auf Englisch als „Nominal Compounds“ oder „Compound Nouns“, im Spanischen meist als „composición sintagmática“ oder „locuciones nominales“ bezeichnet. Ich verwende in dieser Arbeit konsistent *nominale MWE*, in Zitaten in diesem Kapitel können jedoch die genannten oder ähnliche Begriffe auftauchen und sind - je nach Kontext - synonym oder als Teilbereich zu verstehen.

3.1.1 Die spanische Sprache

Das spanische Standardalphabet besteht aus den Buchstaben a bis z, den akzentuierten Vokalen á, é, í, ó, ú, dem Buchstaben ñ sowie dem diakritischen Gebrauch von *ü*, das die Aussprache des *u* in den Sequenzen *gue/gui* markiert (z. B. *pingüino*). Im Hinblick auf diese Arbeit werden MWEs insofern definiert, dass sie ausschließlich mit den Zeichen des spanischen Standardalphabets sowie bestimmten Separatoren (Leerzeichen, Bindestrich) gebildet werden. Diese Beschränkung lässt zwar einige im Spanischen verwendete Entlehnungen, etwa aus dem Katalanischen, Baskischen oder Französischen, außen vor, trägt jedoch insgesamt zu einer deutlicheren Präzisierung der Trefferquote bei.

Wie Denis Maurel ausführt, müssen dennoch einige weitere nicht-standardisierte Zeichen für die Erkennung von MWEs berücksichtigt werden:

> Even though the great majority of elementary forms is restricted to the standard alphabet of the language in question, there are quite a few 'non-standard' word forms involving, in addition to the basic alphabet, a variety of other characters, e.g. numbers, special symbols like the hyphen, or other orthographic means (e.g. symbols not normally occurring inside words, like the dollar, euro, pound and other signs).
>
> [D. Maurel 2005, 38f]

Spanisch (eigenbezogen „Castellano“ oder „Español“) ist eine westromanische Sprache aus der Iberoromania. Neben ihrem Ursprungsort auf der iberischen Halbinsel ist Spanisch in 20 weiteren Staaten Amtssprache, vor allem in Süd- und Zentralamerika sowie der Karibik. Auch in Äquatorialguinea hat es diesen Status. Darüber hinaus gibt es kolonialhistorisch bedingte Sprechergruppen, etwa auf den Philippinen, in Marokko und in der Westsahara. In Nordamerika, insbesondere in den USA, wächst die Bedeutung des Spanischen beständig. Infolgedessen existieren zahlreiche Varietäten des Spanischen, die sich teils erheblich voneinander unterscheiden. Laut dem *Anuario del Instituto Cervantes 2024* übersteigt die Zahl der Sprecherinnen und Sprecher des Spanischen inzwischen 600 Millionen, darunter fast 500 Millionen Muttersprachler und rund 100 Millionen Zweit- und Fremdsprachenlernende.[21]

Der Schwerpunkt liegt für LEMoN aufgrund der Auswahl der Korpora (insbesondere *EuroParl*) auf dem europäischen Standardspanisch. Eine systematische Ausweitung auf andere spanische Varietäten ist, wie [X. Blanco Escoda 2008] argumentiert, für eine umfassende Abdeckung des gesamtspanischen Wortschatzes sicher notwendig und sollte in einem Folgeprojekt Berücksichtigung finden.

> La internacionalisación creciente de todo tipo de actividades económicas, así como la importante presencia de ciudadanos de origen americano en España (consumidores de servicios de proximidad) hace que sea cada día más necesario tener en cuenta las variantes diatópicas en tecnología lingüística.
>
> [X. Blanco Escoda 2008, 1]

[21] Instituto Cervantes (2024): *El español en el mundo. Anuario 2024*. Online verfügbar unter: `https://cvc.cervantes.es/lengua/anuario/anuario_24/` (zuletzt abgerufen am: 10.10.2025).

3.1.2 Spanische NLP-Ressourcen

Obwohl die Lage sich bis 2025 deutlich verbessert hat, war Spanisch in der Korpuslinguistik lange unterrepräsentiert - nicht nur in Bezug auf MWEs, sondern auf sprachtechnologische Ressourcen insgesamt. Bereits 2007 konstatierte Giovanni Parodi im Vorwort zu „Working with Spanish Corpora“:

> Despite the growing interest in the field, studies on Spanish corpus linguistics are scarce, not only in the English-speaking countries but also in Spanish-speaking communities. Spanish is rapidly growing as an international language, making the need for empirical studies of its use more urgent than ever.
>
> [G. Parodi 2007, 1]

Dennoch entspricht die Ressourcenlage auch heute nicht dem Status des Spanischen als zweitmeistgesprochene Sprache der Welt (nach *nativos*), was insbesondere im Bereich der LLMs beklagt wird:

> We launched the #Somos600M Project because the diversity of the languages from LATAM, the Caribbean and Spain needs to be represented in Artificial Intelligence (AI) systems. Despite being the 7.5% of the world population, there is no open dataset to instruction-tune large language models (LLMs), nor a leaderboard to evaluate and compare them.
>
> [M. Grandury 2024, 1]

Aber auch für andere NLP-Bereiche wie Word Sense Disambiguation ([P. Ortega 2024]) oder die Menge an frei verfügbaren Ressourcen wie klinische Fachkorpora ([G. García Subies 2024]) zeigen Artikel, dass die Ressourcenlage weiterhin ausbaufähig ist.

3.1.3 Quantitative Aussagen

Die Größe eines Wortschatzes zu erfassen ist entsprechend schwierig. Das *Diccionario de la lengua española* verzeichnet mehr als 93 000 Lemmata ([Real Academia Española 2014]), die Zahlen von Standardwörterbüchern sind jedoch aus technischer Sicht wenig hilfreich, ein Indiz liefert vielleicht die Größe des *Corpus del Español del Siglo XXI (CORPES XXI)*, das in seiner aktuellen Version 1.3 (Juli 2025) rund 438 Millionen orthographische Formen umfasst

([Real Academia Española 2025a]). Es gibt Schätzungen von Computerlinguisten, etwa für das Englische oder das Französische - und man kann davon ausgehen, dass ähnliche Sprachen insbesondere aus Europa vergleichbare Größenordnungen erreichen. Für das Französische schätzt [D. Maurel 2005, 27], dass die absolute Zahl der *einfachen* lexikalischen Einheiten zwischen ein und zwei Millionen Vollformen liegt. Darauf aufbauend ermittelt er eine Zahl für das gesamte nominale Vokabular von Sprachen, die leicht in den Bereich von zehn Millionen Einheiten steigt:

> Depending on the size of the underlying corpus there are many reasons to believe that just the nominal vocabulary alone of languages with ongoing cultural and scientific communication will easily approach the ten million counts, again if we leave non-technical corpora aside.
>
> [D. Maurel 2005, 28]

Für nominale MWEs liegen keine belastbaren Totalsummen vor. Im varietätenreichen Spanisch ist es zudem schwierig, auch nur zu plausiblen Schätzungen zu gelangen. Timothy Baldwin charakterisiert nominale MWE als die häufigste Art lexikalischer Einheiten (sowohl nach Token- als auch nach Typfrequenz sowie in ihrer Verbreitung über Sprachen hinweg, vgl. [T. Baldwin 2010, 12]). Es handelt sich also um ein Phänomen, dass sehr häufig auftritt, dessen einzelne Realisierungen jedoch mitunter extrem selten sind:

> [. . .] it is common that long terms appear just once even in long corpora. After this appearance, references to this kind of terms often appear only as truncated versions of themselves.
>
> [A. Barrón-Cedeño 2009, 127]

Die enorme Produktivität und der fließende Übergang zwischen idiomatischen und regulären, lexikalisierten und spontanen Bildungen erschweren eine Zählung zusätzlich:

> Uno de los tipos compositivos más productivos del español es la composición sintagmática o composición imperfecta, es decir, el mecanismo de formación de palabras que consiste en la unión de dos o más palabras (como en la composición léxica o prototípica), pero con la particularidad de que no existe una unión gráfica entre sus constituyentes (como en paso de cebra, ojo de buey o perro policía).
>
> [C. Buenafuentes de la Mata 2017, 568]

Für LEMoN genügt daher eine vorsichtige quantitative Einordnung: Nominale MWEs bilden einen äußerst umfangreichen und hochproduktiven Wortschatzbereich, dessen Abdeckung nur korpusgestützt sinnvoll erreichbar ist.

3.2 Taxonomien nominaler MWE

Die im Abschnitt 2.3 dargestellten Kategorisierungsansätze aus der allgemeinen Computerlinguistik einerseits und der Gross-Schule andererseits sind - zumindest für europäische Sprachen - als weitgehend sprachneutral zu betrachten. Ihnen ist gemeinsam, dass sie die Gruppe der nominalen MWEs als einen der zentralen Bestandteile identifizieren. Im Folgenden beschreibe ich Kategorisierungsansätze, die spezifisch nominale MWEs untersuchen und untergliedern. Es existieren auch Ansätze, die das gesamte Feld der MWEs mit Blick auf das Spanische kategorisieren, vgl. etwa [C. Parra Escartín 2018] mit einer systematischen Darstellung aktueller Klassifikationen und einer eigenen belegbasierten Taxonomie sowie [G. Corpas Pastor 1996] mit einer historischen Einordnung der phraseologischen Tradition und früherer Klassifizierungsversuche. Der Fokus des folgenden Überblicks liegt auf der internen Untergliederung der nominalen MWEs.

3.2.1 Nueva gramática de la lengua española (2025)

Die Referenz für die Beschreibung des Spanischen ist die von der Real Academia Española (RAE) in Zusammenarbeit mit der Asociación de Academias de la Lengua Española (ASALE) publizierte *Nueva gramática de la lengua española*. Die jüngste erweiterte Ausgabe von 2025 unterscheidet sich in der Behandlung von MWEs nur marginal von der lange maßgeblichen Auflage von 2011. MWEs erscheinen dort unter der Bezeichnung „unidades léxicas complejas“ und werden als zwischen Lexikon und Grammatik angesiedelte Phänomene beschrieben:

> Las unidades léxicas complejas, como las locuciones o las construcciones con verbos de apoyo, admiten ciertas variaciones sintácticas [...], lo que las coloca en un punto intermedio entre el léxico y la gramática.[22]
>
> [Real Academia Española 2025b, 1.11a]

[22] Komplexe lexikalische Einheiten wie Lokutionen oder Stützverbkonstruktionen lassen bestimmte syntaktische Variationen zu [...], wodurch sie zwischen Lexikon und Grammatik angesiedelt sind. (Übersetzung des Verfassers)

Für den nominalen Bereich unterscheidet die Nueva gramática de la lengua española (NGLE) zwei zentrale Kategorien:

Locuciones nominales: idiomatische, lexikalisierte Ausdrücke, deren Bedeutung nicht aus den Bestandteilen abgeleitet werden kann, z. B. *cabeza de turco* (Sündenbock), *media naranja* (Seelenverwandter).

Composiciones sintagmáticas: Einheiten, die sich morphologisch und syntaktisch wie eine lexikalische Einheit verhalten, deren Bestandteile orthographisch und prosodisch aber unabhängig bleiben, z. B. *pez espada* (Schwertfisch), *casa cuartel* (Kasernengebäude).

Diese Unterscheidung bietet ein klares, aber traditionelles Raster. Die NGLE nimmt keine systematische Klassifizierung nach formalen Mustern[23] (z. B. `NN`, `NA`, `AN`) vor. Damit bleibt ihre Darstellung für korpus- oder computergestützte Arbeiten nur eingeschränkt nutzbar. Gleichwohl markiert die NGLE einen wichtigen Ausgangspunkt: Sie etabliert die Unterscheidung zwischen syntagmatischen Komposita und Lokutionen als Kernkategorien, die in späteren Ansätzen (siehe unten) aufgegriffen, differenziert und für die automatische Erkennung formalisiert werden.

3.2.2 Gloria Corpas Pastor (1996)

Eine wegweisende MWE-Taxonomie für das Spanische stammt aus dem „Manual de Fraseología“ ([G. Corpas Pastor 1996]) der Linguistin Gloria Corpas Pastor. Ihr Werk erschien in einer Phase, in der sich die spanische Phraseologieforschung nach eigener Einschätzung selbst im Weg stand:

> Tan solo se encuentran estudios parciales centrados en la idiomaticidad, o bien clasificaciones incompletas de diverso grado de sofisticación salpicadas, frecuentemente, de una confusa profusión terminológica. Esta situación ha incidido negativamente en la lexicografía, hasta el punto de que, en palabras de Martínez Marín [...], „un tratamiento objetivo y sistemático de la fraseología sigue siendo prágmaticamente imposible hoy por hoy por no haberse realizado los estudios descriptivos“.[24]
>
> [G. Corpas Pastor 1996, 11]

[23] Im weiteren Verlauf dieser Arbeit werden diese Art formaler Muster als *POS-Muster* bezeichnet, die einzelnen Konstituenten als *Main Items*.

[24] Es gibt lediglich Teilstudien mit Fokus aus Idiomatizität, oder unvollständige Klassifikationen in unterschiedlichem Ausarbeitungsgrad, häufig mit einer verwirrenden Terminologie. Diese

Corpas Pastor beschreibt also nicht nur, sondern reagiert mit ihrer Arbeit, die auf bestehenden Klassifikationen (z. B. von Coseriu 1962 und Casares 1992) aufbaut, indem sie aktiv den beschriebenen Mangel zu überwinden versucht: Sie entwickelt eine eigene Systematik. Dafür fasst sie MWEs, die sie als „unidades fraseológicas“ bezeichnet, in zwei Oberkategorien zusammen: „Phrasenebene“ und „Satzebene“.

Für die **Phrasenebene** unterscheidet sie Fixierungsgrade

- auf der Ausdrucksebene (*fijación normal* = *colocaciones*),
- auf der syntaktischen Ebene (*fijación de sistema* = *locuciones*).

Für die **Satzebene** spricht sie von *fijación de habla*, die den Bereich der *enunciados fraseológicos* umfasst.

Jeder dieser Blöcke wird weiter untergliedert. So enthält die Klasse der *locuciones* beispielsweise nominale, adjektivische, adverbiale, verbale, präpositionale, konjunktive und klausale Untergruppen. Für den Kontext dieser Arbeit sind dabei insbesondere die nominalen *Locuciones* und *Colocaciones* relevant, da sie den Kernbereich der in LEMoN modellierten Strukturen bilden.

> Las locuciones nominales están formadas por sintagmas nominales de diversa complejidad. Los patrones sintácticos más productivos son los formados, básicamente, por sustantivo + adjetivo y por sustantivo + preposición + sustantivo.[25]
>
> [G. Corpas Pastor 1996, 94f]

Insgesamt nennt Corpas Pastor die folgenden Typen:

- **NA** (Nomen + Adjektiv)
- **NPN** (Nomen + Präposition + Nomen)
- **NCN** (Nomen + Konjunktion + Nomen)
- **VCV** (Infinitiv + Konjunktion + Infinitiv) („locuciones infinitivas“)

Situation hat sich negativ auf die Lexikographie ausgewirkt, und zwar so weit, dass, in den Worten von Martínez Marín [...], „eine objektive und systematische Behandlung der Phraseologie derzeit pragmatisch unmöglich ist, weil die beschreibenden Studien nicht durchgeführt worden sind“. (Übersetzung des Verfassers)

[25] Nominale Lokutionen bestehen aus Nominalphrasen unterschiedlicher Komplexität. Die produktivsten syntaktischen Muster sind im Wesentlichen Nomen + Adjektiv sowie Nomen + Präposition + Nomen. (Übersetzung des Verfassers)

- **Sonstige** abweichende Muster (wie z. B. cláusulas sustantivadas, expresiones deícticas de otro significado léxico)

Die nominalen Kollokationen beschreibt Corpas Pastor mit Hilfe von Mel'čuks *lexikalischen Funktionen* nicht nur als `NPN`-Muster, sondern auch hinsichtlich ihrer typischen semantischen Relationen:

> Las colocaciones de este tipo indican la unidad de la que forma parte una entidad más pequeña o bien el grupo al que pertenece un determinado individuo [...] El primer sustantivo (el grupo o la unidad) constituye el colocativo, mientras que el segundo es la base (el individuo o la entidad más pequeña).[26]
>
> [G. Corpas Pastor 1996, 74]

Als Beispiele nennt sie *una pastilla de jabón* oder *banco de peces*.

Corpas Pastor verbindet somit eine formale Differenzierung der Nominalmuster mit der funktionalen Perspektive der lexikalischen Funktionen.

3.2.3 Xavier Blanco Escoda (2004)

Xavier Blanco Escoda von der UAB entwirft in [X. Blanco Escoda 2008] eine Typologie spanischer nominaler MWE aus computerlinguistischer Perspektive:

> Las locuciones nominales tienen una importancia fundamental en el procesamiento automático del lenguaje natural, debido a las grandes dificultades de reconocimiento que estas secuencias plantean.
>
> [X. Blanco Escoda 2008, 1]

Theoretisch stützt sich Blanco Escoda explizit auf Mel'čuks Bedeutung↔Text-Modell und unterscheidet (für den nominalen Bereich) vier Typen:

Frasemas nominales completos **(exocéntricos):** Die Gesamtbedeutung enthält keine der Bedeutungen der Konstituenten. Semantisch opak.
z. B. lomo de burro (Bremsschwelle)

[26] Kollokationen dieses Typs bezeichnen entweder die Einheit, zu der ein kleineres Element gehört, oder die Gruppe, der ein bestimmtes Individuum angehört. Das erste Substantiv, also die Gruppe oder Einheit, ist der Kollokator, während das zweite die Basis bildet, also das Individuum oder die kleinere Einheit. (Übersetzung des Verfassers)

***Cuasi-frasemas nominales* (endocéntricos):** Der Gesamtsinn umfasst den Sinn des Kopfnomens und erhält einen spezifischen Mehrwert. Teiltransparent.
z. B. segundo plato (Zweiter Gang)

***Semi-frasemas nominales* (= colocaciones):** Transparente Kombinationen mit lexikalischer Selektion und/oder Restriktion des Modifikators.
z. B. agua con gas (Mineralwasser mit Kohlensäure)

***Pragmatemas nominales*:** Feste Nominalsequenzen aus pragmatischen Gründen (Schilder, Formeln).
z. B. estacionamiento prohibido (Parkverbot)

Neben der semantischen Einteilung dokumentiert Blanco Escoda zusätzlich die *formalen Muster*, unter denen nominale MWEs im Spanischen auftreten:

- **NA** (Nomen + Adjektiv)
- **AN** (Adjektiv + Nomen)
- **NN** (Nomen + Nomen)
- **NPN** (Nomen + Präposition + Nomen)
- **NCN** (Nomen + Konjunktion + Nomen)
- **Residuales** abweichende Muster (wie z. B. Nomen + Präposition + Infinitiv)
- **Sobrecompuestos** (verschachtelte Muster)

Indem Blanco Escoda formale Muster mit einer semantischen Typisierung nach Mel'čuk und spanischspezifischen Besonderheiten verknüpft, gelingt ihm eine Typologie, die sowohl theoretisch differenziert als auch für computerlinguistische Anwendungen unmittelbar verwendbar ist. Der eigentliche Schwerpunkt des Artikels ist die Anlage diatopischer Marker für die verschiedenen spanischen Varietäten.

> En este estudio, nos hemos limitado a dotar nuestro diccionario de locuciones nominales de seis campos complementarios que recogen variantes diatópicas de la entrada utilizadas en Argentina (que abreviaremos Ar), Chile (Ch), México (Me), Perú (Pe), Uruguay (Ur) y Venezuela (Ve). Usaremos la abreviatura Es para España. Próximamente, ampliaremos

> nuestro estudio con los campos correspondientes a la República Dominicana y a Ecuador.
>
> [X. Blanco Escoda 2008, 2]

Darauf aufbauend prüft er, ob funktional äquivalente Ausdrücke zwischen Varietäten unterschiedliche formale Muster realisieren:

> Es frecuente encontrar relaciones de equivalencia entre un NA y un NPN (ya que se trata de las secuencias más numerosas). Con los mismos radicales, tenemos: paso de peatones (Es), paso peatonal (Me) [...]
>
> [X. Blanco Escoda 2008, 4]

3.2.4 Cristina Buenafuentes de la Mata (2021)

Im Routledge Handbook of Spanish Morphology ([C. Buenafuentes de la Mata 2021]) betont Cristina Buenafuentes de la Mata zunächst:

> If the twofold task of defining and delimiting the compound word has occasioned keen discussion in morphological theory [...], establishing a typology of compounds likewise is generally thought to be a highly controversial issue. Thus, it is possible to claim that up to now, there is not yet a universally well-established classification of compounds which has been jointly agreed upon regardless of the theoretical focus applied [...]. Nevertheless, a great many different typologies of compounds based on their inspection from different parameters have enjoyed considerable acceptance in morphological theory.
>
> [C. Buenafuentes de la Mata 2021, 285]

Im Anschluss daran demonstriert sie die Grenzen einer traditionellen Taxonomie der spanischen Nominalkomposita, die lediglich auf den grammatischen Kategorien ihrer Konstituenten basiert, insbesondere eine arbiträr wirkende Mischung aus Part of Speech (POS)-Konstituenten und resultierenden syntaktischen Einheiten darstellt, beispielsweise *lavavajillas* (Geschirrspüler) im Vergleich zu *barbinegro* (schwarzbärtig), die beide Nomina als Konstituenten enthalten, aber selbst einmal ein nominales und einmal ein adjektivisches Kompositum bilden. Aber sie erkennt an:

> Although this classification seems fruitful from a descriptive point of view, as evidenced by the fact that it is used in practically all the descriptions of compounding in Spanish [...], nevertheless, it does not allow accounting for the complexity underlying the compounding processes.
>
> [C. Buenafuentes de la Mata 2021, 286]

Daher schlägt sie vor, diese klassische Taxonomie nicht zu überschreiben, sondern mit einer Supra-Klassifizierung zu versehen, die den zwei Hauptaspekten des *Compoundings* gerecht wird:

- der morphologischen Natur der Konstituenten der resultierenden MWEs
- der intra-kompositorischen Beziehungen, die zwischen den Konstituenten auftreten

Die meisten *Compounds*, die [C. Buenafuentes de la Mata 2021] beschreibt, sind die zusammengeschriebenen spanischen Nominalkomposita, aber ihre Taxonomie und die Erweiterung bezieht sich grundsätzlich auch auf nominale MWEs. Unter den von ihr besprochenen *Compounding Types* fallen lediglich $[\text{N+A}]_{\text{N}}$ und $[\text{N+N}]_{\text{N}}$ compounds in die Kategorie nominaler MWEs, die sie in Hinsicht auf die von ihr beschriebenen Faktoren genauer untersucht.

> To summarize: the classification of compounds proves what Lorenzo (1995, 35) deemed "the shifting sands of compounds". Although it seems practical to count on a compounding typology, on many occasions, the analysis of compounds transcends the boundaries of this categorization; thus, it is necessary to "study every phenomenon independently" (Fábregas 2013, 256). In this sense, the establishment of parallelisms between the special properties evidenced by a compound on an internal level and certain syntactic constructions not only helps to account for the particular behavior of these compounds in Spanish but is also useful for the general characterization of compounding in this language.
>
> [C. Buenafuentes de la Mata 2021, 300]

3.2.5 Vergleich und Bedeutung für LEMoN

Die vier hier herangezogenen Ansätze entstammen verschiedenen Traditionen und Phasen der Entwicklung - entsprechend unterscheiden sich Blickrichtung und Granularität ihrer Taxonomien.

- **NGLE (RAE/ASALE, 2025):** standardgrammatische, weitgehend theorieneutrale Beschreibung zwischen Normierung und Deskription. *Unidades léxicas complejas* werden nach Wortarten (u. a. *locuciones nominales* vs. *composiciones sintagmáticas*) getrennt dargestellt. Der Fokus liegt auf Kategorien und ihren Eigenschaften. Die NGLE klassifiziert primär nach syntaktischer Funktion und idiomatischer Opazität, nicht jedoch nach der formalen Reihenfolge der Konstituenten.

- **Corpas Pastor (1996):** systematisiert die bis dahin terminologisch zerklüftete Phraseologie. Sie ordnet *unidades fraseológicas* zweistufig (Phrasen- vs. Satzebene) und trennt auf Phrasenebene *colocaciones* (fijación normal) von *locuciones* (fijación de sistema). Für den Kontext dieser Arbeit sind die nominalen Unterklassen zentral.

- **Blanco (2004):** arbeitet explizit daten- und anwendungsorientiert und lehnt sich an Mel'čuks Bedeutung↔Text-Modell an. Er unterscheidet nominale *frasemas* (completos, cuasi, semi, sowie *pragmatemas*) und koppelt diese mit spanischtypischen Formmustern (`NA`, `AN`, `NN`, `NPN`, `NCN`, sobrecompuestos) sowie diatopischen Markern.

- **Buenafuentes de la Mata (2021):** zeigt Grenzen rein oberflächengrammatischer Taxonomien auf und ergänzt die traditionelle Kompositionslehre um zwei Querachsen: die morphologische Natur der Konstituenten und die intrakompositorischen Beziehungen zwischen ihnen. Für nominale MWEs sind v. a. $[\mathrm{N}{+}\mathrm{N}]_{\mathrm{N}}$ und $[\mathrm{N}{+}\mathrm{A}]_{\mathrm{N}}$ relevant.

Folgerung für LEMoN: Die Ansätze unterscheiden sich in der Tiefe ihrer Beschreibung und in ihrem Fokus. Sie betrachten jedoch durchweg nicht nur den Aufbau der MWE (mit Ausnahme der NGLE), sondern auch ihre innere Struktur - sei es in der Frage, ob die Bedeutung aus den Bestandteilen ableitbar ist, in der Bestimmung eines Zentrums (Head) oder in der Analyse intra-kompositorischer Beziehungen. Diese Aspekte liegen zwar außerhalb des unmittelbaren Fokus von LEMoN, doch kann das Lexikon als Basis für weiterführende Auszeichnungen in diesen Bereichen dienen.

Trotz unterschiedlicher Zielsetzungen zeigen die Ansätze weitgehend übereinstimmend, dass spanische nominale Komposita (und damit nominale MWEs) auf wenigen, aber hochproduktiven Mustern basieren:

- **NA** (Nomen + Adjektiv)
- **AN** (Adjektiv + Nomen)
- **NN** (Nomen + Nomen)
- **NPN** (Nomen + Präposition + Nomen)
- **NCN** (Nomen + Konjunktion + Nomen)

Genau diese Grundtypen bilden den Kern der in LEMoN implementierten lokalen Grammatiken. Verschachtelte MWEs werden über die *Longest-Matches*-Index-Policy erfasst, die von Blanco genannten Sonderfälle über den Graphen für atypische Muster (vgl. Abschnitt 9).

3.3 Besonderheiten der Morphologie nominaler MWEs

Neben den beschriebenen Strukturmustern spielen bei spanischen nominalen MWEs auch Numerus- und Genus-Kongruenz eine wichtige Rolle. Eine aktuelle detaillierte Analyse der aktuellen und historischen Forschung zu diesem Thema für das Spanische findet sich in der Doktorarbeit von Elia Puertas Ribés von der Universitat Jaume I ([E. Puertas Ribés 2022]). Für diese Arbeit sind diese morphologischen Variationen vor allem deshalb von Relevanz, weil jede Oberflächenform individuell im Korpus auftritt und erkannt werden bzw. in LEMoN kodifiziert werden muss. In Tabelle 1 sind die unterschiedlichen Formen der Realisierung im Einzelnen aufgelistet, vgl. [E. Puertas Ribés 2022].

Typ	Beispiel	Bemerkung
Numerusvariation		
Vollkongruenz	*agujero negro* → *agujeros negros*	Reguläre Pluralisierung: Kopf und Modifikator flektieren synchron.
Teilkongruenz	*casa de campo* → *casas de campo*	Nur der Kopf pluralisiert, abhängiger Bestandteil bleibt im Singular.
Interne Pluralisierung	*orden del día* → *órdenes del día*	Pluralisierung nur auf einem Element der Konstruktion, der Artikel bleibt unverändert. Orthographische Besonderheit: Akzent zur Erhaltung der Betonung.
Nichtkongruente Pluralisierung	*agua mineral* → *aguas minerales* (*el agua* → *las aguas*)	Beide Bestandteile pluralisieren, trotz Massennomen. Artikelwechsel im Plural aufgrund phonologischer Besonderheiten (*el agua fría*, aber *las aguas frías*).
Idiosynkrasien	*gafas de sol*	Pluralia tantum: Singularform unüblich. Andere feste Muster: *tijeras*, *víveres*.
Genusvariation		
Flexible Genuszuordnung	*muerto de hambre* / *muerta de hambre*	Genuswechsel des Kopfes möglich, Bedeutung bleibt identisch.
Epizene / ambige MWE	*oveja negra*, *mar de fondo*	Genus zeigt nicht das natürliche Geschlecht an (epizene) bzw. schwankt in der Tradition zwischen maskulin und feminin (ambig). Interpretation aus Kontext.
Genusfixierung	*balsa de aceite*, *lo de menos*	Festes Genus unabhängig von Konstituenten oder Sprecherintention, teilweise neutrale Strukturen mit *lo*.

Tabelle 1: Numerus- und Genusvariation spanischer nominaler MWEs

Die enge Verzahnung von Morphologie und Syntax im Spanischen verweist auf eine zentrale methodische Herausforderung: Oberflächenformen können nicht losgelöst von syntaktischen Strukturen beschrieben werden. Auch im Kontext von nominalen MWEs zeigt sich, wie in der obigen Tabelle, dass die Grenzen zwischen morphologischer Variation und syntaktischer Regularität fließend sind. Die in Tabelle 1 dargestellten Muster zeigen, dass nominale MWEs stark von Kategorien der Flexionsmorphologie wie Numerus und Genus geprägt sind und dabei häufig mit syntaktischen Regularitäten interagieren - und sich ebenso wie die MWEs, denen sie angehören, mitunter idiomatisch verhalten. Dieses Spannungsfeld ist ein klassisches Beispiel für Morphosyntax im Sinne des „Manual of Romance Morphosyntax and Syntax“:

> In this volume, *morphosyntax* is understood not as the set union of morphology and syntax, but as the interface of grammar in which the components of morphology and syntax interact. [...] Morphosyntax typically makes reference to categories of inflectional morphology, such as person, number, gender, case, tense, aspect, and mood. The flip side of this conception of morphosyntax is that word formation, including compounding and derivation, does not fall within its purview, even if, at times, the boundaries between compositional, derivational and inflectional morphology may appear somewhat blurred.
>
> [A. Dufter 2017, 5]

[E. Puertas Ribés 2022] bietet eine umfassende Zusammenstellung der Morphologie nicht nur nominaler MWEs im Spanischen anhand ihrer morphologischen Variation, formalen Muster und syntaktischen sowie semantischen Funktionen. Ihre Beschreibung verbindet historische Ansätze mit neueren systematischen Analysen und überprüft diese an Korpusdaten. Auch wenn LEMoN keine vollständige Abbildung morphologischer oder intrakompositioneller Prozesse einschließt, sondern sich rein auf die Oberflächenform fokussiert, liefert Puertas Ribés Systematik eine wertvolle Grundlage für weitere lexikographische Arbeiten.

3.4 Determinierer

Unter Determination verstehe ich zum einen die einfache Determination durch klassische Artikel (*el, la, los, las, un, una*) sowie weitere einfache Determinierer (z. B. Demonstrativa, Possessiva), zum anderen komplexe Determinierer.

Unter komplexen Determinierern fasse ich feste oder formelhaft gebundene mehrwortige Determinierer- und Quantor-Ausdrücke wie *todo tipo de*, *una rebanada de* und *un abanico de*. Sie determinieren oder quantifizieren ein folgendes Nomen bzw. eine nominale MWE, ohne selbst Teil ihres lexikalischen Kerns zu sein. Komplexe Determinierer können als Listeme im Wörterbuch hinterlegt sein oder regelhaft, etwa über lokale Grammatiken, erfasst werden.[27]

In [X. Blanco Escoda 2002] unterscheidet Blanco Escoda mehrere Klassen und Unterklassen komplexer Determinierer. Mit Blick auf die Unterklasse der „modifieurs figés“ bemerkt er etwas, das sich verallgemeinern lässt:

> Bien qu'ils soient à recenser comme des collocations par rapport à leurs bases, leur influence sur le fonctionnement de la détermination dans des contextes précis exige une étude d'ensemble de ces unités et leur prise en compte lors de la génération des déterminants de la part d'un système de traitement automatique des langues. [28]
>
> [X. Blanco Escoda 2002, 16]

Auch wenn komplexe Determinierer die Erkennung nominaler MWEs unterstützen, sind sie kein Kerngegenstand dieser Arbeit. Daher beschränke ich mich in dieser Arbeit auf diese knappe Darstellung der Theorie sowie Verweise in den entsprechenden Abschnitten zu Wörterbüchern (7.3.1) und lokalen Grammatiken (9.4.3).

[27] LEMoN macht beides: ein eigenes Lexikon komplexer spanischer Determinierer sowie mehrere lokale Grammatiken.

[28] Obwohl sie in Bezug auf ihre Basen als Kollokationen zu erfassen sind, erfordert ihr Einfluss auf das Funktionieren der Determination in bestimmten Kontexten eine Gesamtuntersuchung dieser Einheiten und ihre Berücksichtigung bei der Generierung von Determinierern durch ein System der automatischen Sprachverarbeitung. (Übersetzung des Verfassers)

4 Akquisitionsverfahren

There are poisons that blind you,
and poisons that open your eyes.

August Strindberg

4.1 Überblick: Verfahrensklassen

In der Literatur begegnen uns für die computergestützte MWE-Erkennung und -Extraktion zahlreiche Bezeichnungen, die ebenso vielfältig sind wie die terminologische Landschaft der MWEs: *Automatic Term Recognition*, *Phrase Detection*, MWE *Identification*, *Keyphrase Indexing*, *Phrase Browsing*, *Phrasing*, MWE *Extraction* oder MWE *Acquisition*

Abhängig von der Art der zu identifizierenden MWEs haben sich unterschiedliche Methoden als nützlich erwiesen. Für diese Arbeit sind vor allem Verfahren relevant, die sich mit nominalen MWEs (in europäischen Sprachen bzw. im Spanischen) beschäftigen. Diese Verfahren lassen sich in fünf Klassen unterteilen: statistische, syntaktische, semantische, kontextbasierte und kombinierte. Ich skizziere sie im Folgenden knapp, ausführlichere Beschreibungen und Vergleiche finden sich etwa in [C. D. Manning 1999, F. Schmidt 2001, E. J. L. Bell 2007, Z. Zhang 2008, H. Heleine 2009, T. Baldwin 2010, C. Ramisch 2015].

Vorweg das Ergebnis: Es gibt bis heute keinen Ansatz zur vollautomatischen, fehlerfreien und umfangreichen Identifikation nominaler MWEs. Entsprechend oft muss eine Abwägung zwischen Kosten und Nutzen getroffen werden (vgl. [C. Ramisch 2012]).

Die Erstellung eines MWE-Lexikons lässt sich sinnvoll in vier Schritte gliedern:

Candidate Extraction / Identification Ermitteln von linguistisch plausiblen MWE-Kandidaten.

Candidate Filtering Filtern der MWE-Kandidaten anhand kontextueller und statistischer Daten.

Candidate Ranking Hierarchisieren der MWE-Kandidaten anhand kontextueller und statistischer Daten.

Candidate Evaluation Manuelle Prüfung der MWE-Kandidaten.

Die konkrete Umsetzung dieser Schritte für diese Arbeit ist im Kapitel zur LEMoN-Pipeline (8) beschrieben.

Da bislang kein vollautomatisches Verfahren existiert, bleibt die manuelle Evaluation unverzichtbar, um die Korrektheit der Ressource zu gewährleisten. In dieser Arbeit liegt der Schwerpunkt auf der systematischen Erzeugung und Filterung von Kandidaten - die abschließende Kuratierung ist als nachgelagerter Prozess vorgesehen.

4.2 Statistische Verfahren

Statistische Verfahren hierarchisieren Kandidaten anhand ihrer distributionellen Eigenschaften und benötigen keine zusätzlichen linguistischen Ressourcen. Aufgrund ihrer Frequenzorientierung erfassen sie seltene oder stark variierende MWEs jedoch nur unzureichend.

Einerseits erlauben sie schnelle Resultate und dienen als exploratives Werkzeug, etwa um Hypothesen zu testen, ohne vorher aufwändige Ressourcenarbeit. Andererseits werden sie in der Praxis oft als Ersatz für notwendige Ressourcenarbeit missverstanden. Das ist problematisch, denn statistische Maße können systematisch kuratierte Lexika nicht ersetzen. Im Kontext von LEMoN spielen rein statistische Verfahren daher keine Rolle.

Ein Grund liegt darin, dass statistische Methoden Texte primär auf Wort- oder N-Gramm-Ebene betrachten. Ohne ergänzende Wörterbücher können sie nicht zwischen einfachen und komplexen lexikalischen Einheiten unterscheiden und verfehlen damit oft die eigentlichen Grundeinheiten der Analyse.

> One reason why for instance many purely statistical approaches to language analysis are not particularly successful is certainly to be found in the simple fact that the units of analysis (i.e. of counting and comparing) are in no way the „interesting“ elements. There is also yet no evidence that all polylexical units can be discovered and then verified with statistical techniques alone; and without such units most statistical investigations will simply address spurious entities (e.g. the individual word forms) most of the time.
>
> [D. Maurel 2005, 17]

Dennoch profitieren auch statistische Analysen erheblich von umfassenden MWE-Wörterbüchern. Zunehmend werden hybride Ansätze verfolgt, die statistische und

linguistische Techniken kombinieren. Technisch sind rein statistische Akquisitionen auf Kollokationen/Kookkurrenzen innerhalb eines lexikalischen Fensters[29] beschränkt. Seltene, variable oder diskontinuierliche MWEs fallen damit systematisch durchs Raster.

Ein verbreitetes Tool ist das `mwetoolkit`[30] (vgl. [C. Ramisch 2010]). Seit 2022 existiert zudem eine Python-Bibliothek[31], die sich in Pipelines integrieren lässt [F. Zagatti 2022].

> The mwetoolkit employs a standard methodology, with candidate extraction followed by candidate filtering, combining Association Measures and descriptive features via a machine learning model; candidates are extracted either as flat n-grams or via morphosyntactic patterns.
>
> [C. Ramisch 2010, 662]

Hypothesentests: Zu den klassischen Hypothesentests zählen t-Test, Chi^2 und Likelihood-Ratio. Alle setzen das gemeinsame Auftreten zweier Wörter ins Verhältnis zu ihren Einzelhäufigkeiten und prüfen die Abweichung von der Nullhypothese (Unabhängigkeit). Der t-Test setzt Normalverteilung voraus, Chi^2 ist robuster gegenüber Verteilungsformen, und der Likelihood-Ratio-Test gilt als zuverlässiger bei kleinen Datenmengen.

Mutual Information: Mutual Information quantifiziert, wie stark das Auftreten eines Wortes Information über das Auftreten eines anderen liefert. Eine bekannte Schwäche ist die Überbewertung seltener Kombinationen. Varianten wie Pointwise Mutual Information oder Enhanced Mutual Information mildern diese Schwäche zwar ab, bleiben aber insbesondere bei kleinen Datenmengen anfällig und schwer vergleichbar.

C-value / NC-value / SNC-value: Der C-Value berücksichtigt eingebettete MWEs. Grundannahme: Kandidaten, die als Teil längerer Einheiten auftreten, sind mit höherer Wahrscheinlichkeit eigenständige Ausdrücke (z. B. *user interface* in

[29] Zahl der nacheinander folgenden Wörter, deren gemeinsame Auftretenshäufigkeit gemessen wird.

[30] `http://mwetoolkit.sourceforge.net/`, zuletzt abgerufen am: 10.10.2025

[31] `https://gitlab.com/mwetoolkit/mwetoolkit3/-/blob/master/mwetoolkit-lib/`, zuletzt abgerufen am: 10.10.2025

graphical user interface). Das in [K. T. Frantzi 2000] beschriebene Verfahren wurde in [A. Barrón-Cedeño 2009] für Spanisch optimiert und kombiniert statistische und linguistische Elemente: POS-basierte Vorfilterung, optional domänenspezifische Stopplisten/Frequenzschwellen und Ranking nach „Termhood“. Die Erweiterungen NC-Value (Kontexteinbezug) und SNC-Value (syntaktische Relationen) erhöhen die Präzision, setzen jedoch zuverlässige POS-Tags bzw. Parser voraus.

Zwischenfazit: Statistische Maße liefern effiziente, aber oberflächliche Signale. Ohne robuste Ressourcen geraten sie bei seltenen oder diskontinuierlichen MWEs schnell an Grenzen. Das motiviert stärkere syntaktische und kontextbasierte Zugänge.

4.3 Syntaxbasierte Verfahren

Syntaktische Verfahren identifizieren oder ordnen Kandidaten auf Grundlage ihrer grammatischen und lexikalischen Eigenschaften und sind dabei auf umfangreiche linguistische Ressourcen angewiesen. Sie gehen über rein statistische Verfahren hinaus, da sie nicht nur Häufigkeiten betrachten, sondern explizit die interne Struktur sprachlicher Einheiten. Beispiele hierfür finden sich in der folgenden Übersicht. Besonders für stark flektierende Sprachen wie das Spanische ist dieser Zugang bedeutsam, da morphosyntaktische Informationen wichtige Hinweise auf die Zugehörigkeit von Einheiten liefern. Syntaktische Verfahren unterscheiden zufällige Wortnachbarschaften besser von stabilen Einheiten, sind aber stark abhängig von der Qualität der zugrunde liegenden Ressourcen und Werkzeuge.

Lemmatisierung: Die Lemmatisierung bzw. Normalisierung gebeugter Formen verändert die Verteilung der Einheiten im Text und ermöglicht präzisere statistische Analysen. Dieser Arbeitsschritt sollte als grundlegende Vorverarbeitung eingesetzt werden.

POS-Tagging: POS-Tagging liefert Informationen über wahrscheinliche Wortarten. So lassen sich gezielt bestimmte Kombinationen suchen oder ausschließen. Für sich genommen ist der Ansatz zu rudimentär, als Vorfilterung jedoch äußerst effizient, da große Mengen irrelevanter Kandidaten frühzeitig ausgesondert werden.

Kongruenz: In stark flektierenden Sprachen ist die Kongruenz ein zentrales Signal für Zusammengehörigkeit. Ein Adjektiv im Singular bezieht sich in der Regel nicht

auf ein Substantiv im Plural. Auch umgekehrt sind stabile Kongruenzbeziehungen ein starkes Indiz für feste Einheiten und damit MWE-Kandidaten.

Linguistisch motivierte Toolsets: Umfassende NLP-Toolkits wie Stanza[32], spaCy[33] oder UDPipe[34] vereinen Tokenisierung, Lemmatisierung, POS-Tagging und Dependency-Parsing (häufig auf Basis der *Universal Dependencies*). Sie können für die MWE-Erkennung genutzt werden, etwa indem syntaktische Relationen Hinweise auf zusammengehörende Einheiten geben. Ihre Genauigkeit hängt jedoch stark von den Modellen und Trainingsdaten ab.

Verbundenheit: Ein Ansatz besteht darin, nicht nur Kookkurrenzen zu betrachten, sondern die direkte syntaktische Abhängigkeit. Seretan [V. Seretan 2011] nutzt das Kriterium der syntaktischen Verbundenheit. Es reduziert zufällige Kookkurrenzen, setzt aber zuverlässige Parser voraus.

Fixiertheit: Viele MWEs zeigen im Vergleich zu freien Wortverbindungen eingeschränkte syntaktische Variabilität (z. B. mangelnde Passivierbarkeit oder interne Modifikationsverbote). Das ist nützlich zur Abgrenzung idiomatischer Einheiten, greift aber nicht universell, da nicht alle MWEs syntaktisch fixiert sind und seltene Vorkommen die Analyse erschweren.

> A[n] [...] approach is to use syntactic fixedness as a means of extracting MWEs, based on the assumption that semantically idiomatic MWEs undergo syntactic variation (e.g., passivization or internal modification) less readily than simple verb-noun combinations [...].
>
> [T. Baldwin 2010, 283]

4.4 Semantikbasierte Verfahren

Semantische Verfahren erstellen oder hierarchisieren Kandidaten anhand ihrer semantischen Eigenschaften. Sie setzen an der Bedeutungsebene an und versuchen, Mehrworteinheiten über inhaltliche Geschlossenheit oder Abweichungen von regulären Bedeutungszuweisungen zu erfassen. Das erfordert leistungsfähige Ressourcen wie semantische Netze, Vektormodelle oder manuell gepflegte Ontologien.

[32] `https://stanfordnlp.github.io/stanza/`, zuletzt abgerufen am: 10.10.2025
[33] `https://spacy.io/`, zuletzt abgerufen am: 10.10.2025
[34] `http://ufal.mff.cuni.cz/udpipe`, zuletzt abgerufen am: 10.10.2025

Nichtaustauschbarkeit: Viele MWEs sind so fixiert, dass ihre Konstituenten nicht durch bedeutungsähnliche Begriffe ersetzt werden können. In [T. Van de Cruys 2007] werden automatisch erzeugte Nomen-Cluster genutzt, um Kombinationen zu identifizieren, die in großen Korpora nie in alternativen Varianten auftreten.

Nichtkompositionalität: Ein weiteres Merkmal ist Nichtkompositionalität. [G. Katz 2006] berechnen Bedeutungsvektoren für idiomatische vs. wörtliche Verwendungen (LSA/LSI) und vergleichen sie per Cosinus-Ähnlichkeit mit den Vektoren der Einzelkonstituenten. Nichtkompositionelle MWEs zeigen signifikant abweichende Werte. Kontextfenster (z. B. 30 Wörter) helfen zusätzlich, die jeweilige Verwendung zu entscheiden. Neuere Arbeiten wie [J. Tanner 2023] nutzen Ressourcen wie WordNet und gleichen Glossare der Konstituenten mit der Gesamtbedeutung ab, um Idiomatizität zu bestimmen.

4.5 Maschinelles Lernen

Verfahren des maschinellen Lernens lernen relevante Eigenschaften aus Beispieldaten statt expliziter Regeln. Unterschieden wird zwischen überwachten (*supervised*) und unüberwachten (*unsupervised*) Verfahren. Die beiden Verfahrenstypen werden zunehmend kombiniert.

Überwachte Verfahren: Basieren auf manuell annotierten Daten, was die Verfahren aufwändig macht. Eingesetzte Verfahren: *Sequence-Labeling*, *Random Forests*, *SVMs*. Sogenannte Weakly supervised Varianten arbeiten mit partiell annotierten Daten (vgl. [K. Kanclerz 2022]).

Unüberwachte Verfahren: Arbeiten mit nicht annotierten Daten und identifizieren Muster entsprechend eigenständig. Der Vorteil ist klar: geringer Bedarf an Vorarbeiten. Dennoch bleibt MWEs zuverlässig ohne umfangreiche Hilfsmittel zu erfassen eine Herausforderung.

Kombinierte Verfahren: *Semi-supervised* Verfahren nutzen annotierte Korpora für manche Sprachen und unannotierte Daten für andere (vgl. [C. Ramisch 2020]). Das erhöht die Ressourcen-Effizienz und erweitert die Sprachabdeckung.

Zwischenfazit: Die Verfahren sind heterogen und nützlich für Teilaufgaben, bieten aber keine umfassende Lösung. Statistische Ansätze sind effizient, aber oberflächlich, syntaktische Methoden sind präziser, aber ressourcenintensiv und semantische Verfahren sind zwar theoretisch reizvoll, aber schwer zu operationalisieren. ML-Modelle sind sehr leistungsfähig, aber intransparent und wenig erklärend.
Für nominale MWEs im Spanischen bleibt das zentrale Problem: Seltene, stark variierende und semantisch schwer fassbare Einheiten entziehen sich einer rein automatischen Erfassung.

4.6 Kontextbasierte Verfahren

Kontextbasierte Verfahren erstellen oder hierarchisieren Kandidaten durch die Kombination lexikalischer, syntaktischer und distributioneller Informationen. Sie können auch als syntagmatische Ansätze begriffen werden:

> Diese syntagmatischen Ansätze beschreiben das Verhalten von Worten und Wortformen in Kontexten und sind damit bezüglich des Umfangs der Bedeutungskodierung nicht festgelegt. Sie können auch als Mittel zur Determinierung der Information in paradigmatischer Hinsicht gesehen werden, da paradigmatische Ansätze ja ohne syntagmatische Information nicht auskommen könnten. Andererseits zeigen die Arbeiten über Kollokationen in großen Korpora, daß Versuche zur statistischen Erfassung von Bedeutung, die auf linguistische Vorabinformation verzichten, die vielschichtigen Regularitäten, denen die Kombinatorik von Lexemen gehorcht, auf problematische Weise vermischen.
>
> [S. Langer 1997, 24]

Einige der unter 4.2 genannten Ansätze beziehen zwar Kontextinformationen ein, bleiben jedoch letztlich rein statistisch und erfassen die komplexen Beziehungen zwischen Kontext und MWE nur unzureichend. Zu den kontextbasierten Ansätzen, die diese Beziehungen nachmodellieren, zählt vor allem die Arbeit mit endlichen Automaten, sogenannten lokalen Grammatiken.

Lokale Grammatiken: Mit lokalen Grammatiken lassen sich sprachliche Phänomene mithilfe gerichteter Graphen und eines Lexikons modellieren. Sie ermöglichen die gezielte Erkennung spezifischer Kontexte. Eingesetzt werden sie u. a. bei der

Identifikation von Eigennamen (vgl. z. B. [M. Geierhos 2006]) oder festgeprägten Ausdrücken in Fachsprachen. Auch zur MWE-Erkennung wurden sie in verschiedenen Sprachen erfolgreich genutzt (vgl. 5). Vor diesem Hintergrund erweisen sie sich als vielversprechender Weg: Lokale Grammatiken erlauben es, linguistische Intuition systematisch zu formalisieren, mit vorhandenen Ressourcen zu kombinieren und zugleich empirisch überprüfbare Ergebnisse zu erzeugen. In dieser Arbeit stellen sie das zentrale Verfahren dar und werden im folgenden Kapitel ausführlich beschrieben.

Teil II

Methoden und Ressourcen

5 Lokale Grammatiken

> *The map is not the territory. The only usefulness of a map depends on similarity of structure between the empirical world and the map.*
>
> Alfred Korzybski

Lokale Grammatiken sind ein Formalismus, der von Maurice Gross in [M. Gross 1993b] vorgeschlagen wurde (bzw. als Konzept bereits in [M. Gross 1989]), um komplexe sprachliche Phänomene präzise und auf eine natürliche, anschauliche und maschinell zu verarbeitende Weise zu repräsentieren. Sie sind in [F. Mallchok 2004, 66] treffend als „language maps“ bezeichnet worden, da sie visuell in Form von Graphen repräsentiert werden. Ein Graph verfügt über jeweils einen Anfangs- und einen Endzustand und verbindet zwischen diesen verschiedene Knotenpunkte mittels gerichteter Kanten. Die Knotenpunkte stellen Filter dar, die etwa Zeichen, Wörter, Wortgruppen oder lexikalische Gruppen beinhalten. Bei einer Textanalyse mithilfe eines Graphen werden nun alle Stellen erkannt und herausgestellt, die vom Anfangszustand einen Pfad über die Knoten zum Endzustand bilden. Das Ergebnis kann in Form einer Konkordanz ausgegeben werden. Unter Konkordanz verstehe ich eine automatisch erzeugte Keyword in Context (KWIC)-Trefferliste: alle Vorkommen eines Musters, jeweils einmal mit linkem und rechtem Kontext, vgl. Abbildung 7.

5.1 Begriffsbestimmung

„Lokale Grammatik“ kann in einem engen und in einem weiten Sinne verstanden werden. In beiden Fällen sind lokale Grammatiken gerichtete Graphennetze, die die Struktur sprachlicher Phänomene nachmodellieren können. Durch die Anbindung an Lexika, den Einsatz von Tokens, Klassen, Operatoren und Kontexten erlauben sie es, nicht bloß einzelne Ausdrücke abzubilden, sondern „funktionale Synonyme“, d. h. Austauschklassen, die sich (bei semantischen Unterschieden) substituieren lassen, ohne den Satz ungrammatisch zu machen, und damit über traditionelle Wortarten hinausgehen:

> Graphen [...] beschreiben ‚funktionale Synonyme‘. Alle von dieser Grammatik beschriebenen Ausdrücke lassen sich in vermutlich jedem englischen Satz gegenseitig austauschen, ohne den Satz ungrammatisch bzw.

> unsinnig zu machen, auch wenn sich die Ausdrücke semantisch unterscheiden (*in the twenties ? in the thirties*). Das unterscheidet die in lokalen Grammatiken beschriebenen Wortklassen von den traditionellen Wortarten. Ein Nomen lässt sich eben nicht durch jedes beliebige Nomen ersetzen.
>
> [S. Nagel 2008, 5]

Das ist die weite Definition lokaler Grammatiken.
Die enge Definition versteht lokale Grammatiken als lexikoneintragartige Beschreibung: Ein Graph erfasst präzise ein einzelnes sprachliches Phänomen - eines pro Graph - und die Gesamtheit dieser Graphen bildet dessen textuelle Vorkommen (einschließlich zulässiger Variation) ab und stellt so die Verbindung zwischen Text und Lexikon her.
In dieser Arbeit folge ich primär der weiten Definition.

5.2 Technischer Rahmen

Lokale Grammatiken entsprechen aus technischer Sicht drei eng verwandten Formalismen:

Finite State Automaton (FSA) Endlicher Automat zur reinen Erkennung.

Recursive Transition Network (RTN) Netzwerk aus FSA-Graphen mit Subgraph-Aufrufen.

Finite State Transducer (FST) Wie FSA, zusätzlich mit Ausgabefunktion (Annotationen, Normalisierung, Satzenderkennung, Substitution). Wird verwendet, wenn Treffer markiert oder transformiert werden sollen.

Für eine ausführliche Analyse von Automaten im Kontext der Sprachverarbeitung siehe [D. Maurel 2005].

Notation und Metasymbole: Lokale Grammatiken werden zu endlichen Automaten kompiliert und auf das Korpus angewendet. Das Resultat ist eine Konkordanz (optional mit Annotationen). Für die Oberfläche genügen wenige Bausteine:

- **Tokens:** Wörtliche Formen im Graphen, z. B. `mujer` oder `Madrid`. Sie matchen genau diese Zeichenfolge. Ist die Zeichenfolge in einem angebundenen

DELA-Wörterbuch vorhanden, wird das Wort entsprechend der Auszeichnung erkannt. Andernfalls wird es als unbekanntes Wort gelistet.

- **Klassen:** In `<spitzen Klammern>` notiert referenzieren sie auf Eigenschaften, die im Lexikon gemäß DELA-Syntax hinterlegt sind (z. B. Wortarten POS, syntaktische Eigenschaften) sowie Gazetteers, z. B. `<N>` (Nomen), `<ADJ>` (Adjektiv), `<CIUDAD>` (Städtenamen).
- **Lexikonbindung:** Klassen können Merkmale wie Genus und Numerus filtern, z. B. `<N.fs>` (Nomen, fem., Sg.). Die Auflösung läuft über die zugrunde liegenden DELA-Wörterbücher.
- **Reguläre Operatoren:** Optionalität `?`, Iteration `+` / `*`, Alternation `|`, Gruppierung `( )`, z. B. `(<N> de)+ <N>`.
- **Subgraph-Aufrufe:** Call-Knoten binden Subgraphen ein und erlauben kontrollierte (endliche) Rekursion für geschachtelte Muster (realisiert durch Doppelpunkt und Graphname, z. B. `:auxi:prep`).

5.3 Bootstrapping

Methodisch greifen beim Entwickeln lokaler Grammatiken drei Arbeitsschritte ineinander: das Beschreiben sprachlicher Strukturen mithilfe lokaler Grammatiken, das Erstellen von Konkordanzen durch Anwendung der Grammatiken auf Korpora (unter Nutzung angebundener Wörterbücher) und das Verbessern der Grammatiken anhand der Analyse dieser Konkordanzen. Diesen iterativen Vorgang bezeichnet man als „Bootstrapping“[35] (vgl. [M. Gross 1999]). Sebastian Nagel fasst einige relevante Punkte prägnant zusammen:

> Charakterisiert ist dieses Verfahren durch den Wechsel zwischen Konkordanzanalyse und Vervollständigung der Grammatik. Nach wie vor ist für beide Schritte die Intuition und Urteilskraft des Linguisten notwendig. [...] Selten ist ein Korpus jedoch so groß, dass es definitiv alle zu beschreibenden Wendungen enthält. In der Folge ist eine Grammatik auch selten vollständig. Die Vervollständigung einer lokalen Grammatik geschieht durch das Hinzufügen neuer Pfade und gestaltet sich damit

[35] Zur Etymologie: „Bootstrapping“ im Sinne von „sich an den eigenen Stiefelschlaufen hochziehen“.

> verglichen mit der Wartung von in anderen Formalismen geschriebenen Grammatiken recht einfach.
>
> [S. Nagel 2008, 10]

Wesentlich ist der Punkt, der die Intuition der Bearbeitenden betont: Es handelt sich um eine halbautomatische Methode, die die Arbeit einer linguistisch geschulten Person voraussetzt, die ein ausreichend großes Korpus nach dem Phänomen durchsucht. Ausgangspunkt ist ein kleiner Graph, der eine für das Zielphänomen typische Struktur am Startknoten beginnend bis zum Zielknoten abbildet. Die folgenden Schritte (vgl. [S. Nagel 2008, 9]) werden anschließend so lange wiederholt, bis keine oder nur noch marginale Verbesserungen erzielt werden:

1) Die lokale Grammatik auf einem geeigneten Korpus anwenden und eine Konkordanz erzeugen.

2) Präzision der Treffer sowie linken und rechten Kontext systematisch nach Mustern durchsuchen, die zur beschriebenen Einheit gehören.

3) Die lokale Grammatik anpassen (Hinzufügen, Editieren oder Entfernen von Kontexten).

4) Mit der überarbeiteten Grammatik erneut eine Konkordanz erzeugen und die aus den Änderungen resultierenden abweichenden Treffer weiter analysieren.

Das Verfahren bleibt bewusst halbautomatisch: Konkordanzen liefern Vorschläge, die Entscheidung über Relevanz und Abstraktionsniveau liegt bei der Linguistin bzw. dem Linguisten. Achtsames Arbeiten ist geboten, denn durch Übereditieren können Ergebnisse verfälscht oder verschlechtert werden. Ein Rückbau kann sich bei komplexen Änderungen als Herausforderung darstellen.

5.4 Schnittstelle Lexikon-Grammatik

Die Lexikon-Grammatik ist der zentrale von Maurice Gross entwickelte Formalismus und verbindet Grammatik und Lexikon eng: Viele grammatische Eigenschaften sind lexikalspezifisch und lassen sich nicht adäquat durch wenige abstrakte Regeln erfassen. Entsprechend werden Einheiten mit idiosynkratischen Eigenschaften als

Listeme explizit gelistet und mit ihren distributionellen, syntaktischen und transformationellen Merkmalen beschrieben [M. Gross 1993b].[36]

Elementare Sätze und Prädikat-Argument-Struktur: Die zentrale Beschreibungseinheit in der Lexikon-Grammatik ist der elementare Satz, also eine basale Prädikat-Argument-Struktur.[37] Transformationen (z. B. Passivierung, Pronominalisierung, Nominalisierung) verknüpfen elementare Sätze mit ihren Varianten. So werden distributionelle und transformationelle Eigenschaften pro Lexem erfasst.

> A lexicon-grammar is constituted of the elementary sentences of a language. Instead of considering words as basic syntactic units to which grammatical information is attached, we use simple sentences (subject-verb-objects) as dictionary entries. Hence, a full dictionary item is a simple sentence with a description of the corresponding distributional and transformational properties.
>
> [M. Gross 1984, 1]

Lexikon-Grammatik-Tabellen: Empirische LG-Tabellen organisieren diese Information: Jede Zeile steht für einen lexikalischen Eintrag (meist Verben), Spalten für elementare Sätze, Merkmale und Konstruktionen (Argumentrahmen, erforderliche Präpositionen, zulässige Transformationen, Selektionsrestriktionen u. a.). In jeder Zelle kann ein Wert kodiert werden (ja/nein/Werte), der angibt, ob der entsprechende Eintrag das Merkmal realisiert. Eine zentrale empirische Beobachtung ist: „keine zwei Einträge teilen exakt dasselbe Profil“. Auf diese Weise belegt die Lexikon-Grammatik empirisch, dass Idiosynkrasie die Regel ist, nicht die Ausnahme [M. Gross 1994].

Nominalphrasen als Argumente: Argumente in den LG-Tabellen werden typischerweise als Nominalphrasen realisiert. Diese reichen von einfachen Formen (`Det+N`) bis zu komplexen und teils feststehenden Einheiten (nominale MWEs). Für die maschinelle Verarbeitung folgt daraus: Sowohl komplexe Determinierer als auch nominale MWEs müssen gelistet oder etwa über lokale Grammatiken erkennbar gemacht werden.

[36] Damit stand die Lexikon-Grammatik in einem deutlichen Kontrast zur generativen Programmatik, vgl. auch die Kritik in [M. Gross 1979].

[37] Für einen Entwurf der Zusammenführung der Grossschen Lexikon-Grammatik mit Mel'čuks Prädikat-Argument-Ansatz vgl. Corpus Calculus [F. Guenthner 2005].

5.5 Mehrworteinheiten und lokale Grammatiken

Nominale MWEs sind zahlreich, einzeln oft selten und zeigen - je nach Fixierungsgrad - Variation wie Flexion oder optionale Modifikatoren. Fachtermini sind ein gut untersuchtes Feld, das sich zu einem hohen Anteil mit nominalen MWEs beschäftigt. In entsprechenden Studien hat sich wiederholt gezeigt, dass die meisten nominalen mehrgliedrigen Termini aus Nomina und Adjektiven (sowie ggf. einer Präposition) bestehen, sowohl im Englischen ([J. S. Justeson 1995]) als auch im Französischen ([B. Daille 2000]), als auch im Spanischen. [R. Estopà Bagot 2001] fassen zusammen, dass die in der Fachterminologie verschiedener Bereiche am häufigsten anzutreffenden nominalen Strukturen `N de N` und `N Adj` sind. Komplexere Muster hingegen (z. B. lange *de*-Ketten) erzeugen deutlich mehr Rauschen, d. h. sind weniger eindeutig allein anhand ihrer POS zu erkennen, weshalb regelbasierte linguistische Systeme an ihre Grenzen stoßen:

> La causa del ruido reside, pues, en la incapacidad de discriminar las UTP [Unidades Terminológicas Poliléxicas, Anm. d. Verf.] a partir de las estructuras morfosintácticas, y este hecho está suscitado por las restricciones de base de las que parten estos sistemas.
>
> [R. Estopà Bagot 2001, 236]

Genau hier setzen lokale Grammatiken an: Durch die Kombination von Kontexten, präziser syntaktischer Modellierung und lexikalischem Wissen lassen sich Kandidaten für nominale MWEs mit kontrollierbarer Balance von Recall und Präzision extrahieren. Der Ansatz ist in mehreren Sprachen produktiv belegt, u. a. Deutsch [F. Guenthner 1998], Französisch [É. Laporte 2008], Italienisch [S. Vietri 2008], Portugiesisch [S. Abalada 2009] und Koreanisch [J. Han 2018].

> Lokale Grammatiken für die Erkennung der textuellen Vorkommen von Phrasemen können das Bindeglied zwischen Text und Lexikon bilden. Die Grammatik modelliert auf abstrakte Weise die Menge der textuellen Vorkommen jedes einzelnen Phrasems.
>
> [C. Kunze 2007, 309]

Zur Veranschaulichung folgt im nächsten Abschnitt eine kompakte Beispielgrammatik. Die konkreten, für diese Arbeit entwickelten Grammatiken zur Generierung der Kandidatenlisten spanischer nominaler MWEs werden in Kapitel 9 ausführlich beschrieben.

5.6 Beispiel einer lokalen Grammatik

Ein einfaches Beispiel soll diesen Formalismus verdeutlichen. Der in Abbildung 6 dargestellte Graph ist auch für Laien gut lesbar: Dem Startzustand links folgt entweder das Wort *a*, *de* oder *en*, gefolgt von einer Stadt. Bei `<CIUDAD>` handelt es sich um ein Metazeichen, das durch die spitzen Klammern angezeigt wird. Im Lexikon, auf das dieser Graph zugreift, sind Städtenamen mit dem Zusatz `CIUDAD` markiert und damit einer nur auf Städte bezogenen Suchanfrage zugänglich. Metazeichen lassen sich in den verwendeten lexikalischen Ressourcen frei definieren und ermöglichen sowohl syntaktische als auch semantische oder pragmatische Auszeichnungen. Die Abbildung 7 zeigt einen Auszug aus der Konkordanz dieses Beispielgraphen, mit dem der Text „Trafalgar“ von Benito Pérez Galdós durchsucht wurde[38]. Auffällig ist, dass sogar Städtenamen, die aus mehr als einem Wort bestehen, durch die einfache Suche nach `<CIUDAD>` gefunden wurden - mit Ausnahme von Medina-Sidonia, das offenbar nicht im Lexikon als Stadt ausgezeichnet war und daher nicht vollständig erkannt wurde.

Da *a*, *de* und *en* der Wortart Präposition zugeordnet werden können, wäre auch eine Suche nach `<PREP>` anstelle der drei Einzelwörter in Frage gekommen. Dann wären jedoch auch alle anderen Präpositionen erfasst worden, die vor dem Namen einer Stadt stehen können, also zum Beispiel *entre*, *hasta* oder *por*. Je nach Zielsetzung lassen sich lokale Grammatiken so erstellen, dass sie mittels funktionaler Synonyme nicht nur in der Lage sind, alle gültigen Varianten einer Äußerung zu erkennen, sondern sie auch zu produzieren.

[38] Der Text liegt Unitex als Beipielkorpus für Spanisch bei.

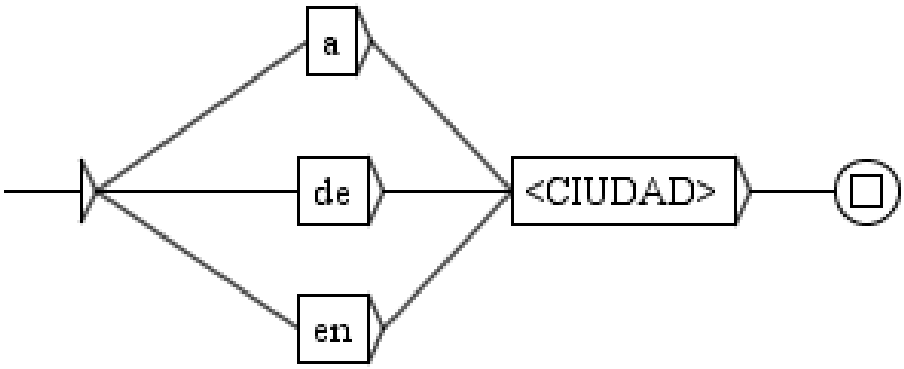

Abbildung 6: Beispiel einer lokalen Grammatik

21 21 por la noche supimos en Cádiz el éxito del combate.{ S} Lo dic
Aquel coloso, construido en La Habana con las más ricas maderas c
ba por la ramas, y una vez en La Habana gastó 10 000 duros en ciert
na! —Pues cuando yo estuve en Madrid el año último —prosiguió el em
pues sólo estaban de paso en Medina-Sidonia.{ S} Mis ángeles tutela
fondeaban frente a Cádiz o en San Fernando.{ S} Como nunca pude sati

Abbildung 7: Beispiel einer Konkordanz

6 DELA-Wörterbücher

> *The lexicon is like a prison - it contains only the lawless, and the only thing that its inmates have in common is lawlessness.*
>
> Anna-Maria Di Sciullo und Edwin Williams

6.1 Begriffsbestimmung

In diesem Kapitel kläre ich Terminologie und Rahmen für das Lexikondesign. Das „Léxico de Expresiones Multipalabra Nominales“, kurz LEMoN, ist das im Rahmen dieser Arbeit erstellte, maschinenlesbare Lexikon spanischer nominaler MWEs im Dictionnaire Electronique du LADL (DELA)-Format. Die Wahl des Formats und die zugrunde liegenden Prinzipien begründe ich hier.

Dass MWEs Listeme sind, also grundlegende bedeutungstragende Einheiten, habe ich in Kapitel 2 dargelegt. In an Menschen gerichteten Wörterbüchern (Leserwörterbüchern) werden MWEs dennoch meist nur sporadisch und unsystematisch erfasst. Digitalisierte Varianten solcher Leserwörterbücher werden als Machine Readable Dictionary (MRD) bezeichnet, also lesbar aber nicht ohne weiteres verarbeitbar. Für die Sprachverarbeitung hingegen werden sogenannte Machine Traceable Dictionary (MTD) erstellt und verwendet, also elektronische Wörterbücher, deren Inhalte so explizit und formal kodiert sind, dass Programme zuverlässig darauf zugreifen können. In dieser Arbeit verwende ich die Bezeichnungen „elektronisches Wörterbuch“ und „elektronisches Lexikon“ synonym, beide beziehen sich auf MTD.

6.2 Technischer Rahmen

6.2.1 Datenmodelle

Die Gestaltung von MTDs beginnt mit der Entscheidung für ein Datenmodell, also die Organisation der lexikalischen Daten innerhalb der Lexikondatei. Relevant sind u.a. Verarbeitungsgeschwindigkeit, Dateigröße sowie Les- und Editierbarkeit. Im Folgenden skizziere ich vier verbreitete Modelle mit ihren Stärken und Schwächen.

Lemmabasiert Der lemmabasierte Ansatz ist gewissermaßen der „natürliche Ansatz“ eines MTD. Die Grundeinheiten des Lexikons sind damit die einzelnen

Lemmata, denen nun Eigenschaften wie POS-Tags und Regeln zur Flexion zugeordnet werden können. Diese Form der Lexikongestaltung hat den Vorteil, dass sie keine unnötige Redundanz beinhaltet. Um ein lemmabasiertes MTD etwa zum Taggen eines Textes effizient verwenden zu können, kann man die nur durch Flexionsregeln repräsentierten flektierten Formen kompilieren und etwa in Form eines *mixed* oder wortformenbasierten Lexikons (siehe unten) speichern (vgl. [É. Laporte 2005b]).

Gemischt Ein gemischtes Lexikon-Modell ist eine Variante des lemmabasierten. Der Unterschied ist, dass die flektierten Formen nicht in Form von Regeln repräsentiert sind, sondern assoziiert zum jeweiligen Lemma explizit aufgelistet werden. Nachteilhaft ist in diesem Falle vor allem die zum Teil starke Redundanz, in Bezug auf Wortstämme beträgt sie ca. 1 zu 10, in Bezug auf die Flexionsmuster gar 1 zu 1.000[39]. Des Weiteren ist das gemischte Modell nur bedingt anwendbar für Sprachen, die entweder keine Flexion kennen, etwa Chinesisch, oder für agglutinierende Sprachen, die nahezu unzählbar viele Varianten eines Wortes produzieren können, etwa Türkisch (vgl. [É. Laporte 2005b]).

Wortformbasiert Im Gegensatz zu den beiden lemmabasierten Ansätzen werden beim wortformbasierten Ansatz die Lemmata den Vollformen zugeordnet und nicht umgekehrt. Nachteilig erweist sich hier, neben den Nachteilen des gemischten Modells, dass eine hinzugefügte flektierte Form keine Auskunft über evtl. weitere fehlende Formen desselben Lemmas gibt (auch hier vgl. [É. Laporte 2005b]).

Indexbasiert Indexbasierte MTD sind zur möglichst effizienten Abfrage von Lexika als endliche Automaten oder lexikalische Bäume organisiert. Sie entkoppeln die Laufzeit von der Lexikongröße. Für ein standardisiertes Austauschformat eignen sie sich jedoch nur bedingt, da erstens die Struktur stark von Kompressions-/Lookup-Verfahren abhängt und zweitens interaktive Aktualisierung erschwert ist. In flektierenden Sprachen ist es daher üblich, gemischte oder wortformbasierte Lexika in ein indexbasiertes Arbeitsformat zu kompilieren und so Updates effizient auszurollen (vgl. [É. Laporte 2005b, 4]).

Für LEMoN nutze ich ein gemischtes Arbeitsformat. Die Kompilierung in ein indexbasiertes Format erfolgt erst für den Einsatz in Tools.

[39]Die errechneten Verhältnisse beziehen sich auf ein Lexikon, in dem MWEs nicht explizit beachtet wurden.

6.2.2 Gestaltungsanforderungen

Die gestellten Anforderungen an ein MTD sind inhaltlicher, technischer und formaler Natur. Neben der unmittelbaren Gewährleistung des beabsichtigten Nutzens zielen die Anforderungen auch auf eine nachhaltige Nutzbarkeit, Wartbarkeit und Wiederverwendbarkeit der Daten ab.

Grundlegende Gedanken zur Gestalt von elektronischen Wörterbüchern und den Anforderungen, die an sie gestellt werden müssen, wurden schon früh von [B. Courtois 1990, 3] formuliert. Dementsprechend müssen MTD vor allem die Kriterien „Vollständigkeit“, „Explizitheit“ und „maschinelle Verarbeitbarkeit“ erfüllen.

In [F. Guenthner 1994] werden vor allem die Kriterien Vollständigkeit der gewählten Teilmenge, Korrektheit der Ableitung und Abgeschlossenheit der Regularitäten betont. Zusätzlich wird jedoch auch auf einige damals noch nicht besonders verbreitete und eher pragmatische Anforderungen verwiesen: Portabilität, Wiederverwendbarkeit und Verwendung von Standards.

Weitere Anforderungen, die an MTD gestellt werden sollten, sind zum einen Theorieneutralität, um das Lexikon einer Vielzahl von Anwendungen und Herangehensweisen, die zum Zeitpunkt der Erstellung möglicherweise noch gar nicht bekannt sind, zu erschließen, und zum anderen Menschenlesbarkeit, die den Anwender des Lexikons von dessen Ersteller unabhängig macht (zu letzterem vgl. [C. Kunze 2007, 109]).

> If we aim at a reusable multifunctional lexicon, obviously our choices, for each parameter (such as size, depth of information, levels of linguistic description, explicitness, etc.), should be guided by the most demanding, the widest, and the deepest requirements. However, these should be corrected by a careful analysis of a) the possibility of actually achieving these most general requirements, in accordance with the state-of-the-art (in linguistic theory, software technology, etc.), and b) the cost- effectiveness of such ambitious prospects as opposed to more modest but more realistic and economically viable strategies.
>
> [N. Calzolari 1991, 269] (nach [S. Langer 1997])

6.2.3 Kriterienkatalog

Für die praktische Umsetzung nominaler MWEs in einem elektronischen Lexikon - also ihre systematische Erfassung, Kategorisierung und Kodierung - sind insbeson-

dere folgende Fragen zentral (vgl. [F. Guenthner 2004]):

1) Welche nominalen MWEs müssen in Wörterbüchern beschrieben werden?

2) In welcher Form sollen MWEs in Wörterbüchern beschrieben werden?

3) In welche Kategorien sollen diese MWEs im Wörterbuch eingeteilt werden?

4) Welche Informationen sollen mit den MWEs in den einzelnen Kategorien kodiert werden?

Aus diesen Fragen resultiert der folgende Kriterienkatalog. Die Kriterien helfen zu entscheiden, welche Daten erfasst werden und wie die sprachlichen Informationen im Lexikon kodiert werden sollen, auch wenn es nicht immer möglich ist, alle Anforderungen gleichzeitig zu erfüllen.

C1 Vollständigkeit (VOL) Das abzudeckende Inventar des Lexikons muss zuvor bestimmt und dann (möglichst) vollständig erfasst sein. Dies ist aufgrund der Produktivität von Sprache natürlich nur für bestimmte Bereiche überhaupt denkbar.

C2 Explizitheit (EXP) Jede Information im Lexikon muss explizit kodiert sein. Was nicht explizit kodiert ist, muss vollständig und korrekt hergeleitet werden können (siehe Korrektheit).

C3 Maschinenlesbarkeit (MSL) Die Informationen müssen formal stringent kodiert sein, um eine Verarbeitung durch Software zu ermöglichen.

C4 Korrektheit (KOR) Das Regelwerk und die Auszeichnung zur Herleitung von flektierten Formen aus Grundformen und umgekehrt muss korrekt sein.

C5 Abgeschlossenheit (ABS) Alle morphologischen und semantischen Regularitäten, die für das abgedeckte Vokabular relevant sind, müssen vorhanden sein.

C6 Portabilität (POR) Das MTD soll in verschiedenen Anwendungen und Implementierungsformen eingesetzt werden können.

C7 Wiederverwendbarkeit (WIE) Das MTD soll nicht nur für einmalige Experimente angelegt werden, sondern darüber hinaus wiederverwendet werden können.

C8 Standardunterstützung (STU) Die Nutzung von internationalen Standards kann die langfristige Verwendbarkeit eines Lexikons und dessen Interoperabilität mit anderen Anwendungen ermöglichen.

C9 Theorieneutralität (THE) Das MTD soll dergestalt konzeptioniert werden, dass es nicht nur im Licht einer einzelnen linguistischen Theorie verwendbar ist.

C10 Menschenlesbarkeit (MEL) Das MTD soll so ausgezeichnet und gespeichert sein, dass es nicht nur von Maschinen verwertbar und von Programmiererinnen/Programmierern erweiterbar ist, sondern auch von fachkundigen Laien eingesehen und bearbeitet werden kann.

Operative Lesart für LEMoN: Für diese Arbeit lese ich die Kriterien wie folgt praktisch ein:

- **C1 Vollständigkeit:** Ich priorisiere hohe Abdeckung in den Trainingsdaten (Recall) gegenüber strenger Ausfilterung (Precision). Ein zweifelhafter Kandidat im Lexikon ist weniger problematisch als ein fehlender gültiger Ausdruck.
- **C2 Explizitheit:** Jede gelistete Einheit ist eine Mehrworteinheit. Die Bestandteile erhalten grobe POS-Tags. Ambiguitäten werden nicht „weganno-tiert“, sondern als getrennte Einträge sichtbar gehalten.
- **C3 Maschinenlesbarkeit:** Die Auszeichnung erfolgt in DELA.
- **C6/C7 Portabilität/Wiederverwendbarkeit:** Die Ressource ist nicht an eine einzelne Anwendung gebunden. Varianten stehen als Daten, nicht bloß als implizite Regeln.
- **C9 Theorieneutralität:** Abgesehen von der Grundannahme „MWEs sind Listeme“ bleibt die Auszeichnung minimalistisch und theorieoffen.

Im folgenden Kapitel beschreibe ich das von mir verwendete DELA-Format und referenziere dabei die hier angegebenen Kriterien. In Tabelle 4 ist die Abgleichmatrix dargestellt, die die Kriterienerfüllung von DELA gegenüber diesem Katalog prüft.

6.3 Das DELA-Format

Im Folgenden beschreibe ich das von mir verwendete DELA-Format ausführlich, das seinen Ursprung ebenso wie lokale Grammatiken im LADL hat.

DELA ist der Standard in der Gross-Schule und lässt sich in andere Formate verlustarm transformieren, solange alle relevanten Informationen abbildbar sind. Für zahlreiche Sprachen existieren umfangreiche DELA-Lexika, etwa:

- Deutsch [F. Guenthner 1994]
- Englisch [A. Monceaux 1995]
- Französisch [B. Courtois 1990]
- Griechisch [M. Constant 2003]
- Italienisch [S. Vietri 2008]
- Katalanisch [J. Sastre Alaiz 2007]
- Koreanisch [C. Hwang 2018]
- Malagasy [J. N. A. Ranaivoarison 2015]
- Portugiesisch [E. Ranchhod 1999]
- Russisch [S. Nagel 2006]
- Spanisch [X. Blanco Escoda 2001]

Das DELA-System ist als Wörterbuchfamilie gegliedert, deren Mitglieder die folgenden sind:

Mitglieder der Wörterbuchfamilie:

DELAS (*mots simples*): einfache Lemmata

DELAC (*mots composés*): komplexe Lemmata (MWEs)

DELAF (*formes fléchies*): einfache Vollformen mit Flexionsmerkmalen

DELACF (*composés avec formes fléchies*): komplexe Vollformen mit Flexionsmerkmalen

6.3.1 Struktur

Makrostruktur: Ein DELA-Wörterbuch ist eine einfache Textdatei. Die Makrostruktur folgt einem gemischten Modell: ein Eintrag pro Zeile. In DELAF/DELACF stehen flektierte Formen zusammen mit Lemma und Kategorien. DELAS/DELAC erfassen Lemmata bzw. komplexe Lemmata ohne explizite Flexionsangaben.

Mikrostruktur und Felder: Die Mikrostruktur basiert (in der Vollform-Variante) auf bis zu fünf Feldern in fester Reihenfolge. Grob lässt sie sich in drei Segmente gliedern: (1) beobachtete flektierte Form (Vollform), (2) Grundform (Lemma) und (3) anschließende Annotationen. Je nach Lexikon und Wortart/Lemma werden alle oder nur einige der Felder verwendet. Die Syntax unterscheidet sich zwischen nichtflektierten (DELAS/DELAC) und flektierten (DELAF/DELACF) Lexika nur darin, dass bei ersteren die kanonische Form und der Punkt entfallen. Flexionsmerkmale können in Form von Grammatiken notiert werden, um eine gewisse Flexionsfunktionalität zu gewährleisten. Die fünf Felder sind:

Flektierte Form (Vollform) Obligatorisch, gefolgt von einem Komma.

Kanonische Form (Lemma) Optional. Im Spanischen ist dies in der Regel das Maskulinum Singular, aber bei ausschließlich femininen Lemmata das Femininum Singular. Für Verben wird der Infinitiv verwendet.

Grammatische und/oder semantische Kategorien Jedes Lemma benötigt wenigstens eine grammatische oder semantische Kategorie, die durch einen Punkt von der kanonischen Form abgetrennt wird. Mehrere Kategorien werden mit + akkumuliert.

Flexionsinformationen Optional, nur in Lexika für flektierte Formen (DELAF/DELACF). Es können eine oder mehrere Flexionsmerkmale (d. h. Genus und Numerus sowie Deklination und Konjugation) beschrieben werden, die Trennung erfolgt dabei durch Doppelpunkte, die nach OR-Semantik verstanden werden.

Kommentar für den menschlichen Bearbeiter Optional, wird durch einen Schrägstrich eingeleitet.

Knapp zusammengefasst:

- `DELAF/DELACF: Vollform, Lemma.Kat:Flex / Kommentar`
- `DELAS: Lemma.Kat / Kommentar` (ohne Vollform, daher kein Komma)
- `DELAC: Komplexe Form, .Kat / Kommentar` (leeres Lemma, Punkt bleibt)

Die Mitglieder der DELA-Familie folgen derselben Mikrostruktur (Vollform, Lemma, Kategorien, Flexionsangaben, Kommentar), die Zeilensyntax unterscheidet sich jedoch leicht: Bei DELAS entfällt die Vollform (daher kein Komma). Bei DELAC steht nach der komplexen Vollform ein Komma, dann `.Kategorien` - der Punkt signalisiert das leere Lemma explizit (Lemmafeld entfällt, der Punkt nicht). In Codebeispielen verwende ich die Abkürzungen `Kat` = (grammatische/semantische) Kategorien und `Flex` = Flexionsangaben. Die Felder werden über verschiedene standardisierte Trennzeichen voneinander abgegrenzt (siehe Tabelle 2).

Tabelle 2: Trenn- und Markierzeichen in der DELA-Zeilensyntax

Zeichen	**trennt/markiert**	**Bemerkung / Beispiel**
`,`	Vollform - Lemma	`manzanas, manzana.N:fp /...`
`.`	Lemma - Kategorien	`manzana.N`, bei DELAC: `KomplexeForm, .N` (leeres Lemma)
`+`	Kategorienakkumulation	`N+Conc` (mehrere Kategorien für denselben Eintrag)
`:`	Kategorien - Flexionsangaben	`N:fp`, mehrere `:`-Segmente werden als Alternativen interpretiert (OR), z. B. `grande, grande.A:ms:fs`
`/`	Eintrag - Kommentar	`...:fp /Äpfel` (freier Text, optional)

Weitere Metazeichen: `=` innerhalb von Tokens steht für Bindestrich/Leerzeichen-Varianz.

6.3.2 Behandlung von Mehrworteinheiten

> Isolating a simple word from its context, as parsing programs do in a first step, can be quite misleading. [...]
> As a consequence, dictionaries of compound words (i.e. generalized idioms) must be constructed for computer analysis [...]
>
> [M. Gross 1999, 1]

Ziel dieser Arbeit ist, MWEs explizit im Lexikon zu listen. Wie u. a. beschrieben in [F. Guenthner 2004], sollten einfache und komplexe Lemmata im elektronischen Wörterbuch gleichförmig kodiert werden. Dennoch erfordert ihre Kodierung darüber hinaus einige Designentscheidungen, weil sich ihr Verhalten unter Umständen von einfachen Lemmata unterscheidet. In diesem Abschnitt skizziere ich die Eignung und den Einsatz des DELA-Formats für die Kodierung von MWEs. Die konkreten Lexika, die ich für diese Arbeit erstellt und verwendet habe, sind in Abschnitt 7 aufgeführt. Da das DELA-Auszeichnungsformat aus der Gross-Schule hervorgeht, ist es grundsätzlich MWE-tauglich. Im Folgenden diskutiere ich die Behandlung nominaler MWEs im DELA-Format entlang eines natürlichen Bogens: Vom Zuschnitt dessen, was überhaupt als MWE erfasst wird, über die Notation eines Eintrags, den Umgang mit Varianten und Flexion sowie die Frage, wie mit konkurrierenden Lesarten umzugehen ist, bis hin zu Codes/Metadaten und den Grenzen des Formats einschließlich projektspezifischer Aufnahmekriterien. Wo es zum Verständnis beiträgt, vermerke ich die entsprechenden Entscheidungen für LEMoN. Punktuelle Verweise auf den Kriterienkatalog (C1-C10) markieren den Bezug, die Gesamtbewertung fasst Tabelle 4 zusammen. Wo Entscheidungen zum Aufnahmeumfang oder zur Kodierung nötig sind, treffe ich sie projektspezifisch (vgl. Erfassungsbereich für die Auswahlpolitik und Grenzen des Formats für die Abbildung nicht-kontinuierlicher Muster). Die Bewertung erfolgt entlang des Kriterienkatalogs (Tabelle 4).

Erfassungsbereich: Der Gegenstand dieser Arbeit ist die lexikalische Erfassung nominaler MWEs. Die Entscheidung für ein typbezogenes Lexikon ist pragmatisch sinnvoll, vgl. [P. Mogorrón Huerta 2004, 391]. In LEMoN werden MWEs bewusst inklusiv gefasst: Neben idiomatischen Einheiten werden auch Kollokationen und lexikalisierte, regulär gebildete Sequenzen aufgenommen. Benannte Entitäten (Named Entities) sind eingeschlossen. Leitlinie ist ein Recall-orientierter Ansatz: Eine reguläre Wortgruppe im Lexikon ist weniger problematisch als eine fehlende tatsächliche

MWE, die Aussortierung kann ggf. in nachgelagerten Schritten geschehen (C1, C5). Die Repräsentation erfolgt oberflächenformbasiert in DELAF. Erfasst werden die im Korpus attestierten Vollformen in DELA-konformer Zeilensyntax (C3, C6, C7).

Einträge und Ambiguität: Komplexe Einheiten werden weitgehend analog zu einfachen Formen notiert. Varianten, Homonyme/Polysemie-Fälle und Ambiguitäten lassen sich ohne Zusatzformate abbilden (C2, C10). Die DELA-Syntax erlaubt mehrere Einträge für dieselbe Oberflächenform. Unterschiedliche Lesarten (z. B. syntaktische Muster oder semantische Klassen) koexistieren, bis sie außerhalb des Lexikons, etwa durch Kontexte, disambiguiert werden.

Varianten: Oberflächenvarianten (Bindestrich/Leerzeichen, optionaler Artikel etc.) werden explizit aufgeführt. Bei Bedarf können sie über dieselbe kanonische Form zusammengeführt werden. Das Metazeichen = (z. B. `años=luz`) markiert Bindestrich-/Leerzeichen-Varianz und unterstützt eine robuste Korpuserkennung (vgl. Tabelle 2, C2, C6). Pluralbedingte Akzentverschiebungen (z. B. `orden` → `órdenes`) werden als eigene Vollformen erfasst (C2).

Flexion: Für die Flexion komplexer Einheiten gibt es zwei Wege: Erstens die explizite Auflistung beobachteter Formen und zweitens regelbasierte Generierung. Wo Flexionsmuster kodiert sind, lässt sich zwar aus DELAS/DELAC automatisch ein DELAF/DELACF erzeugen. Bei MWEs ist dies jedoch nicht trivial: Häufig flektiert nur der Kopf, andere Bestandteile bleiben fix. Teils bestehen interne Kongruenzen (Genus/Numerus) oder optionale Einschübe. In dieser Arbeit liste ich beobachtete Formen explizit, um die Trefferquote in heterogenen Korpora zu maximieren. Eine Normalisierung/Reduktion kann nachgelagert erfolgen. Den erhöhten Speicher- und Pflegeaufwand nehme ich zugunsten robuster Erkennung bewusst in Kauf (C1, C3, C7).

Auszeichnung und Erweiterbarkeit: DELA erfasst morpho-syntaktische Eigenschaften über kompakte Codes direkt im Eintrag (vgl. Tabelle 3). Der etablierte Kanon deckt Standardfälle ab, projektspezifische Codes lassen sich ohne Änderung der Zeilensyntax ergänzen. So bleiben die Einträge explizit und das DELA-Format theorieoffen (C2, C9).

Tabelle 3: Grammatische und Flexionscodes

Grammatische Codes		**Flexionscodes**	
Code	**Beschreibung**	**Code**	**Beschreibung**
A	Adjektiv	m	maskulin
ADV	Adverb	f	feminin
COMDET	Komplexe Determinierer	n	neutrum
CONJ	Konjunktion	s	singular
DET	Determinierer	p	plural
N	Nomen		
PREP	Präposition		
PRON	Pronomen		
V	Verb		

Grenzen des Formats & Abgrenzung: DELA erfasst kontinuierliche Sequenzen. Varianten mit Dislokation oder optionalen Einschüben werden als separate Einträge (Vollformen) erfasst und als Kandidaten im Lexikon aufgenommen. So bleiben die Einträge flach und wartbar (C7, C10), während kompositionelle Effekte außerhalb des Lexikons behandelt werden (C3, C4).

6.3.3 Abgleich mit dem Kriterienkatalog

Zusammenfassend ergibt sich für die Behandlung von MWEs in DELA bzw. in LEMoN ein konsistentes Gesamtbild: Die getroffenen Entscheidungen zielen auf Explizitheit, Menschen- und Maschinenlesbarkeit und sichern zugleich Portabilität sowie Wiederverwendbarkeit (C2, C3, C6, C7, C10). Innerhalb der Möglichkeiten von DELA bleibt das Lexikon theorieoffen (C9). Der bewusst gewählte Recall-orientierte Ansatz akzeptiert dabei eine gewisse Redundanz zugunsten robuster Erkennung, ggf. nachgelagerte Normalisierungen sind möglich. Eine zusammenfassende Bewertung anhand des oben angegebenen Kriterienkatalogs findet sich in Tabelle 4.

Tabelle 4: Abgleichmatrix: DELA und Kriterienkatalog (C1–C10)

ID	Kriterium	Status	Kurzbegründung
C1 (VOL)	Vollständigkeit	ermöglicht	DELA begrenzt den inhaltlichen Umfang nicht. Als Format erlaubt es vollständige Abdeckung, garantiert sie aber nicht.
C2 (EXP)	Explizitheit	ja	Klare Mikrostruktur (Vollform, Lemma, Kategorien, Flexion, Kommentar). Informationen stehen explizit in der Zeile.
C3 (MSL)	Maschinenlesbarkeit	ja	Zeilenorientiertes, strikt formales Textformat. DELA ist problemlos pars- und konvertierbar.
C4 (KOR)	Korrektheit	ermöglicht	Korrektheit ist Ressourceneigenschaft. DELA unterstützt korrekte Ableitungen via kodierten Merkmalen, garantiert sie aber nicht.
C5 (ABS)	Abgeschlossenheit	ermöglicht	Morphologische/semantische Regularitäten können als Codes/Grammatiken erfasst werden. Vollständigkeit der Regeln liegt außerhalb des Formats.
C6 (POR)	Portabilität	ja	Einfache Textdateien, bewährte Toolkette (z. B. Kompilierung in indexbasierte Lexika) und verlustarme Konvertierbarkeit.
C7 (WIE)	Wiederverwendbarkeit	ja	Stabil, menschen- und maschinenlesbar. Für unterschiedliche Pipelines und Zwecke wiederverwendbar.
C8 (STU)	Standardunterstützung	teilweise	Standard in der Gross-Schule. Keine formale ISO-Norm, aber gut in andere Standards (z. B. LMF) transformierbar.
C9 (THE)	Theorieneutralität	ja	Offene, erweiterbare Codes. Keine Bindung an eine spezifische Theorie.
C10 (MEL)	Menschenlesbarkeit	ja	Kompakte, flache Syntax. Einträge sind ohne Spezialsoftware les- und editierbar.

Status-Legende: ja = erfüllt; teilweise = erfüllt mit Einschränkungen; ermöglicht = vom Format unterstützt, aber nicht garantiert (ressourcenabhängig).

6.3.4 Beispielhafte Lexikoneinträge:

Eine beispielhafte Auswahl von Lexikoneinträgen für einfache lexikalische Einheiten im DELA-Format findet sich in Tabelle 5, die verschiedenen Codes werden in Tabelle 3 aufgelöst.

Tabelle 5: Felder und Beispiele in der DELA-Syntax

	Flektierte Form	Kanonische Form	Kategorien	Flexions-angaben	Kommentar
1	ordenador,	ordenador.	N+Conc	:ms	/Einfaches Lemma (Substantiv, Sg.)
2	manzanas,	manzana.	N+Conc	:fp	/Substantiv im Plural, kanonische Form angegeben
3	grandote,	grandote.	A	:ms	/Adjektiv, augmentierte Form
4	encontraste,	encontrar.	V	:2sJ	/2. Sg. Indefinido (Prät.)
5	manzana de la discordia,	manzana de la discordia.	N	:fs	/MWE
6	años=luz,	año=luz.	N	:mp	/Plural, Leerzeichen-Varianz
7	La Habana,	Habana.	N+Prop	:fs	/Artikelvariante zum Lemma *Habana*
8	primeros auxilios,	primeros auxilios.	N	:mp	/Pluralia tantum
9	órdenes del día,	orden del día.	N	:mp	/Plural mit Akzentwechsel am Kopf

6.3.5 Alternative Auszeichnungsformate

Es gibt immer wieder Versuche, neue Standards für die Arbeit mit elektronischen Lexika (und überhaupt mit linguistischen Ressourcen) zu etablieren, beispielsweise das auf eXtended Markup Language (XML) basierende Lexicon Markup Frame-

work (LMF)[40] [ISO/TC 37/SC 4 2006]. Dieses wurde vom International Standards Organization (ISO)-Gremium „Technical Committee 37 - Terminology and other language and content resources“ standardisiert, vgl. [C. Kunze 2007, 121ff] LMF ist eine Synthese gemeinsamer Schwerpunkte verschiedener, zuvor gängiger Lexikonmodelle wie dem OLIF-Modell. Besonders hervorheben möchte ich den Standard „OntoLex-Lemon“: Der Beiname dieses Standards ist nur zufällig namensgleich mit dem Titel dieser Arbeit. LEMoN ist ein Akronym für das hier entwickelte Wörterbuch.

Ein kurzer Überblick ohne Anspruch auf Vollständigkeit:

- **LMF (ISO 24613):** Metamodell für Austausch/Interoperabilität. `https://www.lexicalmarkupframework.org/`, zuletzt abgerufen am: 10.10.2025.
- **TEI Lex-0:** Schlankes TEI-Profil für Lexika. `https://github.com/TEIC/TEI-Lex0`, zuletzt abgerufen am: 10.10.2025.
- **OntoLex-Lemon (W3C):** Verknüpfung von Lexika und Ontologien. `https://www.w3.org/community/ontolex/`, zuletzt abgerufen am: 10.10.2025.
- **OLIF/TBX:** Austausch in Terminologie/TMS-Workflows (industriell). `https://www.tbxinfo.net/`, zuletzt abgerufen am: 10.10.2025.

Für die vorliegende Studie ist DELA das prädestinierte Format. Ein späteres Mapping (z. B. nach LMF oder OntoLex-Lemon) ist möglich, aber in der Regel nicht erforderlich.

6.4 Entwicklungsprozess (Überblick)

Die Erstellung eines elektronischen Lexikons ist nach [S. Langer 1998] in zwei Arbeitsschritte untergliedert:

(1) Segmentierung der Lemmata (hier: nominale MWEs) mittels lokaler Grammatiken aus den Zielkorpora (Details zu den Grammatiken in Kapitel 9 und zu den verwendeten Korpora in Kapitel 7).

(2) Zuordnung der für das Lexikon benötigten lexikalischen Informationen (insb. POS, Numerus, Genus).

[40] `http://www.lexicalmarkupframework.org/` (zuletzt abgerufen am: 10.10.2025)

Technisch gesehen lassen sich beide Schritte vereinen, indem die Grammatiken, die zur Erkennung der MWEs verwendet werden, als FST implementiert werden.

Auf diesem Wege lässt sich automatisch die Information, aus welchen Wortarten die MWE zusammengesetzt ist, auszeichnen - vorausgesetzt, die POS sind im zugrundeliegenden Wörterbuch korrekt und vollständig hinterlegt. Ein in seiner Wichtigkeit nicht zu unterschätzender dritter Schritt ist die manuelle Kontrolle. Da es kein automatisches Verfahren gibt, um zuverlässig MWEs zu identifizieren, ist es ebenfalls unmöglich, automatisiert die Korrektheit identifizierter Kandidaten zu bestimmen. Ebenso wenig lässt sich zuverlässig automatisch ermitteln, ob ein POS korrekt gesetzt wurde. POS-Tags werden für die Bestandteile automatisch zugewiesen und bleiben bis zur expliziten Auflösung ambig. Auch komplexe Flexionsmuster können regelbasiert erzeugt werden (z. B. mit Multiflex, [A. Savary 2009]), in dieser Arbeit liste ich jedoch explizit gelistete Vollformen und lagere entsprechende Schritte nach.

Die Pipeline erzeugt Kandidaten, führt identische Formen zusammen und zählt Vorkommen für die Auswertung. Eine sprachliche Normalisierung oder zusätzliche Kodierung findet nicht statt. Der vollständigen Beschreibung der Pipeline ist Kapitel 8 gewidmet. Die Evaluation von Wörterbuch und Pipeline findet sich in Kapitel 10. Überblicksartig lässt sich die Pipeline folgendermaßen darstellen:

(1) Kandidatenerzeugung: Lokale Grammatiken laufen als Transduktoren über die Zielkorpora und geben Oberflächenformen nominaler MWEs aus. Das Vorgehen ist recall-orientiert (C1) und hält Ambiguitäten sichtbar (C2).

(2) Aggregation & Frequenz: Identische Oberflächenformen werden zusammengeführt. Vorkommen werden gezählt. Es erfolgt keine weitere Normalisierung und keine zusätzliche Anreicherung. Ambiguitäten in Form von alternativen POS-Strukturen bleiben bestehen (C2, C10).

(3) Export ins Arbeitsformat: Die aggregierte Liste wird unverändert, skriptbasiert in eine DELA-Datei überführt. Alternativ bleiben die Konkordanzen mit Kontexten für Qualitätskontrolle, Optimierung der lokalen Grammatiken oder weitere Analysen erhalten. Auszeichnung als syntaktische Klasse `N` ist für nominale MWEs standardmäßig gesetzt, als weitere Information ist die Zahl und Art der Main Items (MIs) gespeichert. Der Export ist reproduzierbar, d. h. identische Ein-

gaben führen mit denselben Skripten/Parametern zu identischen Ausgaben (C3, C6, C7).

Überblicksartig:

- Kandidatenerzeugung mittels lokaler Grammatiken (Transduktoren),
- Aggregation identischer Oberflächenformen & Frequenzzählung,
- Export ins DELA-Arbeitsformat (deterministisch, ohne inhaltliche Änderungen).

7 Daten und Werkzeuge

We shape our tools and thereafter our tools shape us. ...
We shaped the alphabet and it shaped us.

John M. Culkin

7.1 Unitex

Die zentrale Komponente dieser Arbeit ist Unitex/GramLab[41], im Folgenden kurz Unitex. Es handelt sich um eine mehrsprachige Open-Source-Suite zur Korpusverarbeitung, die drei Ressourcentypen zusammenführt: Korpora, elektronische Wörterbücher und lokale Grammatiken. Unitex bildet damit den Rahmen für die im Anschluss beschriebenen Korpora und Lexika. Die lokalen Grammatiken werden in Kapitel 9 beschrieben.

In Funktionalität und Handhabung knüpft Unitex an die am LADL entwickelte Software INTEX[42] und an die Tradition der Lexikon-Grammatik an. Unitex stellt eine Graphical User Interface (GUI) zur Entwicklung lokaler Grammatiken sowie deren Anwendung auf Korpora bereit. Korpora können vorverarbeitet und getaggt werden. Unitex arbeitet mit Ressourcen, die am LADL zunächst für das Französische entwickelt wurden, insbesondere mit elektronischen Wörterbüchern im DELA-Format, vgl. Kapitel 6.3. Der in dieser Arbeit verwendete Ansatz zur Extraktion spanischer MWEs basiert auf mit Unitex entwickelten lokalen Grammatiken und nutzt vorhandene DELA-Ressourcen, was die Reproduzierbarkeit der Experimente sicherstellt.

Grundkonfiguration In allen Experimenten wurden die folgenden Standardeinstellungen verwendet. Laufspezifische Abweichungen oder Ergänzungen sind jeweils an den betreffenden Stellen im Kapitel 10 dokumentiert.

Allgemeine Einstellungen:

- Encoding: UTF-16LE, Zeilenende LF
- Satzsegmentierung: übernommen von [C. Beiras Cunqueiro 2005]

[41] Unitex/GramLab v3.3, `https://unitexgramlab.org/`, zuletzt abgerufen am 10.10.2025.
[42] entwickelt von Max Silberztein.

- Wörterbücher: Alle Default-Lexika werden angewendet (siehe Tabelle 8)

UnitexToolLogger *Locate*:

- Graph: `start.grf`
- Index: Longest Matches
- Search limitation: Index all occurrences in text
- Search algorithm: Paumier (2003)
- Grammar outputs: Merge with input text

UnitexToolLogger *Concord*:

- Context length left: 0
- Context length right: 0
- Sort according to: Center, Left

7.2 Korpora

Dieser Abschnitt beschreibt die eingesetzten Korpora und ihre Rolle in den drei Arbeitsphasen. Die Tabellen dokumentieren Herkunft und Eigenschaften der Korpora. Im Fließtext erläutere ich, wofür die Korpora jeweils genutzt wurden, welche Alternativen geprüft wurden und warum bestimmte Ressourcen verworfen oder bevorzugt wurden. So werden Datengrundlage, Einsatz und Entscheidungen transparent.

Für die Entwicklung nutze ich ein mehrstufiges Vorgehen, das sich am Zusammenspiel von Test- und Validierungskorpora orientiert.

7.2.1 Korpora in der Testphase

Zunächst habe ich ein kompaktes *El Mundo*-Zeitungskorpus zur Erstellung und Anpassung der lokalen Grammatiken im Bootstrapping-Verfahren verwendet. Es ist groß genug, um relevante Strukturen zu enthalten, zugleich aber klein genug, um auch bei vielen Iterationen kurze Ladezeiten zu gewährleisten. Um Überanpassung an einen einzelnen Datensatz zu vermeiden, habe ich die Entwicklung schrittweise

auf größere, heterogene Ressourcen ausgeweitet und systematisch geprüft. Verwendung fanden insbesondere Ausschnitte größerer Ressourcen wie der *Wikipedia* sowie weitere privat erstellte Webkorpora.

Im Fokus standen dabei das Verständnis des Verhaltens und Vorkommens spanischer nominaler MWEs, eine stabile Erkennung, die Vermeidung von Übererkennung, die Definition relevanter Kontexte, die Implementierung komplexer Determinierer sowie die Trefferquote (Recall) bei Domänenwechseln. Diese Phase diente dem Aufbau der LEMoN-Grammatiken, ersten Robustheitsprüfungen unter unterschiedlichen Domänenbedingungen und der rudimentären Ausgestaltung der Pipeline.

7.2.2 Korpora in der Validierungsphase

Die in dieser Phase eingesetzten Ressourcen sind wesentlich umfangreicher. Die Arbeit an den LEMoN-Grammatiken dient in dieser Phase dazu, Annahmen aus der Testphase an größeren und weniger fokussierten Datensätzen zu bestätigen oder zu falsifizieren, die Möglichkeiten und Grenzen des verfolgten Ansatzes auszuloten und die Pipeline gezielt auszugestalten. Ein iterativer Wechsel zwischen Test- und Validierungsphase sowie den jeweils genutzten Ressourcen liegt in der Natur des Bootstrapping-Verfahrens. Für die Validierung habe ich auf verschiedene Quellen zurückgegriffen, darunter das *MultiUN*-Korpus, einen vollständigen Dump der spanischen *Wikipedia* von 2011 sowie weitere private Webkorpora.

In der Validierungsphase habe ich intensive Explorationen und frühe prototypische Hauptläufe durchgeführt. Mehrere Ansätze wurden verworfen. Ungeeignet erwies sich insbesondere die Kopplung des großen *El País*-Zeitungskorpus mit einer XML-Pipeline, Transducer-Outputs mit umfangreicher XML-Annotation und einer Datenbankanbindung. Dieser Aufbau erhöhte die Komplexität, band Ressourcen in Technik und Wartung und verschob den Fokus weg von den linguistisch motivierten Untersuchungszielen. Daraus folgte die Entscheidung, die Validierung ressourcenschonend zu gestalten und die Pipeline auf ein klar abgegrenztes Trainingskorpus sowie dateibasierte, reproduzierbare Schritte zu fokussieren.

Es existieren zudem eine intern erstellte Taxonomie semantischer Klassen auf Basis von [S. Langer 1997] und [X. Blanco Escoda 2010] sowie darauf aufbauende lokale Grammatiken und Wörterbücher. Der Einsatz semantischer Klassen wurde für LEMoN intensiv exploriert, letztlich jedoch nicht in die Pipeline übernommen. In einer thematisch fokussierten Vorstudie (vgl. [D. Stein 2012]) habe ich den Einsatz semantischer Klassen zur Extraktion nominaler MWEs systematisch erprobt und als

nicht zielführend bewertet.

Die LEMoN-Grammatiken haben am Ende dieser Phase ihre finale Form angenommen. Nach Versuchen mit hochgradig spezialisierten Grammatiken ist ein systematisch gedachter Ansatz entstanden, der sich über mehrere Korpora bewährt hat.

7.2.3 Korpus für den Hauptlauf

Nach Abschluss der Validierungsphase führe ich den Hauptlauf durch, aus dem das Wörterbuch `LEMoN.dic` hervorgeht. Ich wende die finalen LEMoN-Grammatiken in der erarbeiteten LEMoN-Pipeline auf das Zielkorpus *EuroParl* an. Es enthält Transkriptionen von Parlamentsdebatten des Europäischen Parlaments, sowohl spanische Originaltexte als auch Übersetzungen aus anderen Amtssprachen der EU. Die Wahl fiel auf *EuroParl*, weil es im Unterschied zum ursprünglich vorgesehenen *El País*-Korpus thematisch klar abgegrenzt und moderat dimensioniert ist und damit vollständige Verarbeitung, belastbare Reproduzierbarkeit und manuelle Sichtungen in ausreichender Breite ermöglicht.

Die Verwendung von *EuroParl* bringt Vorteile und Einschränkungen mit sich, die bei der Interpretation der Ergebnisse berücksichtigt werden müssen.

Vorteile:

- Paralleles Korpus in allen EU-Amtssprachen mit Potenzial für sprachübergreifende MWE-Analysen.
- Textsorte Parlamentsdebatten unterscheidet sich deutlich von den in der Testphase genutzten Zeitungs- und Enzyklopädietexten und erlaubt damit die Prüfung der Übertragbarkeit auf eine neue Domäne.
- Moderate Größe ermöglicht manuelle Sichtungen der Ergebnisse.

Einschränkungen:

- Hoher Übersetzungsanteil mit möglichen Effekten von „translationese“ wie stilistischen Interferenzen und vereinfachten Strukturen, vgl. [Y. Graham 2019].
- Dominanz des iberischen Spanisch, Varietätenmischung durch Übersetzungen jedoch nicht ausgeschlossen.
- Hochstilisiertes Register mit fachsprachlich geprägten Topoi und geringer Nähe zur Alltagssprache.

- Verschriftlichte gesprochene Sprache mit Satzabbrüchen, Selbstkorrekturen, Interjektionen und gelegentlich fehlerhafter Satzsegmentierung.

Die Übersicht zu Rolle und Inhalt steht in Tabelle 7, die Größen sind in Tabelle 6 dargestellt.

Tabelle 6: Größen der Korpora in UTF-16LE (Bytes = Dateigröße, Zeichen = Unicode-Zeichen nach Dekodierung, Token = grobe Wortzählung via regulärem Ausdruck `\w+`)

Korpus	Bytes	Zeichen	Tokens
El Mundo	85 323 488	42 661 744	6 731 526
El País	8 385 858 132	4 192 929 066	695 600 933
EuroParl	1 382 375 752	691 187 876	280 130 088
MultiUN	5 311 757 326	2 662 883 777	408 473 545
Wikipedia.es	8 618 373 678	4 309 186 839	1 679 064 883

Tabelle 7: Eingesetzte Korpora

Name	Rolle	Varietät	Domäne	Zeitraum
El Mundo	Entwicklung	Spanisch (Spanien)	Zeitungstexte	Jul–Okt 1995
El País	Validierung	Spanisch (Spanien)	Zeitungstexte	1976–2009
Wikipedia.es (Dump)	Validierung	Spanisch (international)	Enzyklopädie	12.07.2011 (Dump)
EuroParl (ES)	Hauptlauf	Spanisch (Spanien)	Politik (Parlamentsdebatten)	1996–2011
MultiUN (ES)	Validierung	Spanisch (international)	Politik (UN-Dokumente)	2000–2009

Korpora und Lizenzen:

- *EuroParl v7 (ES)* frei für Forschung, vgl. [P. Koehn 2005].
- *Wikipedia.es Dump 2011* CC-BY-SA, Abruf 12.07.2011.
- *MultiUN (ES)* frei nutzbar, vgl. [A. Eisele 2010].
- *El País* und *El Mundo* intern erstellt, nicht öffentlich.

7.3 Wörterbücher

LEMoN nutzt lokale Grammatiken und benötigt dafür möglichst vollständige und konsistent ausgezeichnete Lexika im DELA-Format.

Neue Korpora enthalten in der Regel Wörter, die in keinem vorhandenen Lexikon verzeichnet sind, sogenannte Out of Vocabulary (OOV)-Einheiten. Diese müssen zunächst ins Lexikon aufgenommen und annotiert (d.h. lexikographiert) werden, um ihre Erkennung zu ermöglichen und gegebenenfalls Lesarten gezielt zu steuern. Lexikographierung ist also ein Prozess, in dem zweierlei entschieden wird: Erstens, ob ein OOV-Wort in das Lexikon überführt werden soll (was z.B. bei fremdsprachlichen Wörtern, Schreibfehlern oder Kodierungsartefakten nicht immer der Fall ist) und zweitens, wie, d.h. mit welcher Annotation und in welchem Wörterbuch es aufgenommen werden soll. Das Lexikon `DELAF-for-lemon.dic` beinhaltet unter anderem die von mir angelegten OOV-Listen. Tabelle 8 stellt die verwendeten Wörterbücher zusammen, diese sind in Unitex als *default dictionary* gekennzeichnet und werden somit bei jedem Lauf verwendet. Die Vorbereitung von Korpora inklusive der Lexikographierung von OOV-Wörtern ist in Abschnitt 8.1 beschrieben.

7.3.1 Verwendete Wörterbücher

Die in LEMoN verwendeten Wörterbücher sind in drei Klassen gegliedert: Basislexika, Speziallexika und technische Lexika. Tabelle 8 führt sie mit Kernangaben auf. Im Folgenden skizziere ich Funktion und Umfang.

- Basis
  - `DELAF`
    Beschreibung: Hauptressource für einfache lexikalische Einheiten. Es ist das mit Unitex gelieferte `DELAF` (vgl. [X. Blanco Escoda 2001]).
  - `DELAF-for-lemon`
    Beschreibung: Selbst erstelltes Ergänzungslexikon. Zur vereinfachten Pflege wurden zuvor separat geführte thematische Wörterbücher und OOV-Listen konsolidiert.
  - `basic-override-`
    Beschreibung: Priorisiertes Override-Wörterbuch (kennzeichnendes Minus am Dateiende). Einträge werden exklusiv in der hier kodierten Lesart

interpretiert. Alternative Lesarten aus anderen Wörterbüchern werden ignoriert.

- Spezial
  - `eswiki-categories`, `eswiki-titles`
    Beschreibung: Grundinventar an MWEs aus der spanischen Wikipedia. Erstellt aus Artikeltiteln (`eswiki-titles.dic`) und Artikelkategorien (`eswiki-categories.dic`).
  - `ncomp_uab`
    Beschreibung: Externes Lexikon für MWEs und Komposita der UAB (`ncomp_uab.dic`). Nutzung mit freundlicher Genehmigung. Die Einträge entsprechen nominalen MWEs im Sinne von LEMoN.
  - `personas`
    Beschreibung: Selbst erstelltes Gazetteer mit Personennamen. Als `N` ausgezeichnet und dadurch Teil der MWE-Erkennung.
  - `sem_GEO`
    Beschreibung: Selbst erstelltes Gazetteer mit geographischen Bezeichnungen. Als `N` ausgezeichnet und dadurch Teil der MWE-Erkennung.
- Technisch
  - `complex-det`
    Beschreibung: Wörterbuch polylexikalischer Determinierer[43].
  - `romnum`
    Beschreibung: Liste römischer Zahlen.
  - `syn_prefix`
    Beschreibung: Liste von Präfixen, als Einzelwörter erfasst, um gelegentliche Getrenntschreibung abzufangen.

[43] Zu komplexen Determinierern vgl. Abschnitt 3.4.

Tabelle 8: Verwendete DELA-Wörterbücher (gesetzt als „default dictionaries“ in Unitex)

Name	**Einträge**	**Funktion**	**Quelle**
Basis			
DELAF	638 000	Basis	Unitex
DELAF-for-lemon	541 314	Basis	selbst erstellt
eswiki-categories	66 202	MWE	Wikipedia-Extraktion
eswiki-titles	246 987	MWE	Wikipedia-Extraktion
ncomp_uab	90 650	MWE	extern (UAB)
Spezial			
personas	10 411	Personennamen	selbst erstellt
sem_GEO	77 995	Geographische Namen	selbst erstellt
Technisch			
basic-override-	175	Disambiguierung	selbst erstellt
complex-det	1 973	Determinierer	selbst erstellt
romnum	1 001	Römische Zahlen	selbst erstellt
syn_prefix	84	Präfixe	selbst erstellt

7.3.2 Wörterbuchabdeckung auf EuroParl

EuroParl ist das Korpus des Hauptlaufs. Für die Entwicklung der LEMoN-Grammatiken wurde es nicht verwendet. Eine punktuelle Vorverwendung einzelner Ressourcen lässt sich nicht ausschließen. Lexikalisch ist *EuroParl* jedoch nicht systematisch erschlossen und trifft die LEMoN-Grammatiken daher weitgehend unbekannt.

Zur Abdeckung der in diesem Kapitel beschriebenen Default-Lexika: Es verbleiben **53 092** nicht erkannte Wortformen (OOV). Bezogen auf die Tokenzahl des Korpus ergibt das

$$\frac{53\,092}{280\,130\,088} \cdot 10^5 \;=\; 18{,}95$$

also rund *19 OOV je 100 000 Token* ($\approx 0{,}019\,\%$). Diese Zahl dient hier als Abdeckungsindikator für die Ressourcen. Die Einordnung erfolgt im Hauptlauf, vgl. Abschnitt 10.2.

Teil III

LEMoN

8 LEMoN-Pipeline

> *A genius is a man who takes the lemons that Fate hands him and starts a lemonade-stand with them.*
>
> Elbert Hubbard

In Kapitel 4 habe ich die vier Standardschritte der MWE-Akquisition benannt: Extraktion, Filterung, Ranking und Evaluation. Im Folgenden bilde ich diese auf die in dieser Arbeit eingesetzte LEMoN-Pipeline ab:

Extraction Transduktion der lokalen Grammatiken über die vorverarbeiteten Korpora mit den Default-Lexika mittels Unitex. Ausgabe sind Oberflächenformen nominaler MWEs.

Filtering Filter über die im Graph modellierten Kontexte (u. a. Kongruenz, Determinierer/Präpositionen) sowie technische Bereinigungsschritte (OOV-/Artefakt-Handling).

Ranking Aggregation identischer Oberflächenformen und Frequenzzählung pro Datei/Korpus und gesamt.

Evaluation Nachgelagerte manuelle lexikographische Prüfung der exportierten DELA-Einträge (`LEMoN.dic`).

Damit ist der Abgleich mit den Verfahrensklassen aus Kapitel 4 hergestellt. Nun führe ich die Schritte entlang der in Abbildung 8 skizzierten Pipeline aus, beginnend mit der Korpusaufbereitung.

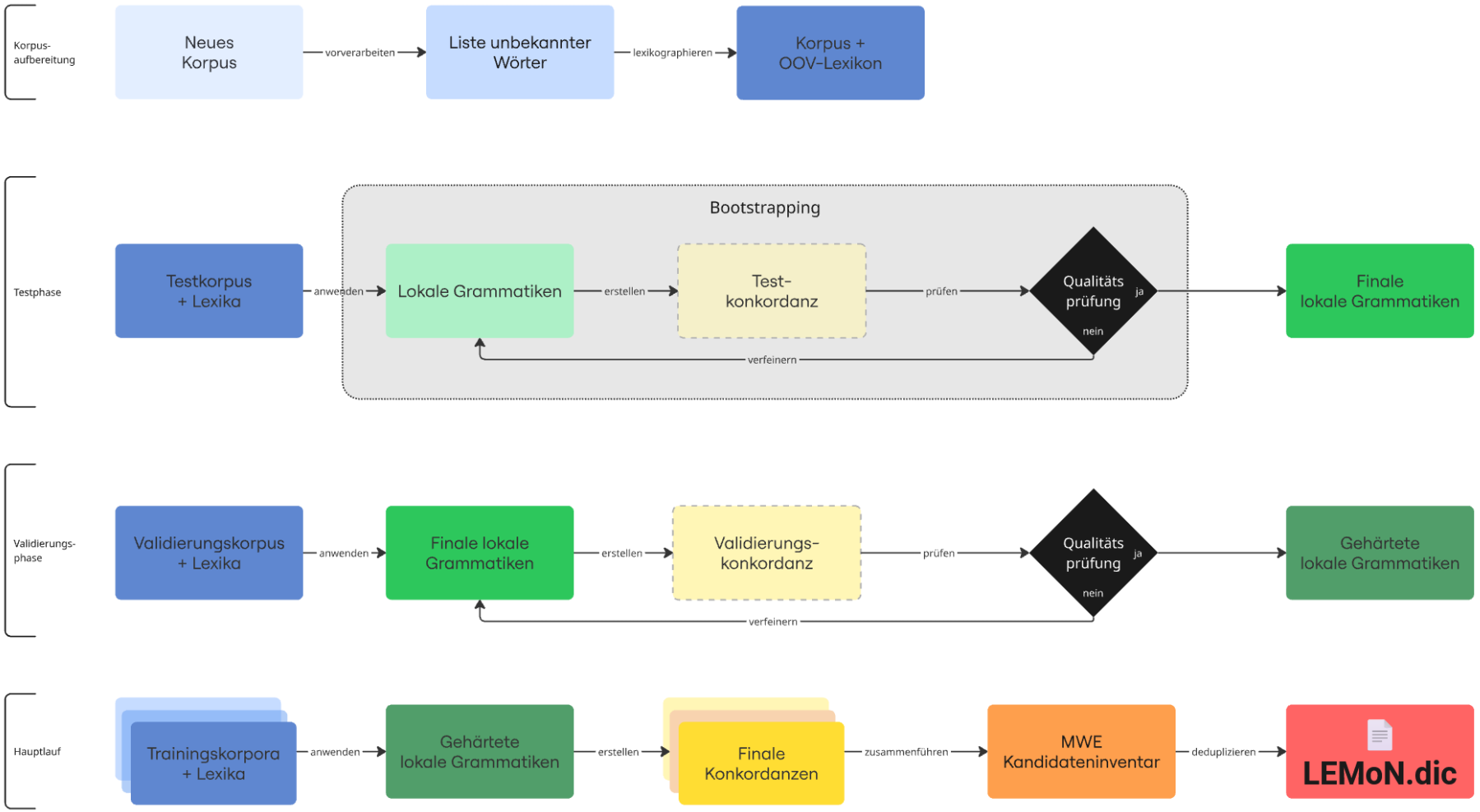

Abbildung 8: Die LEMoN-Pipeline

8.1 Korpusaufbereitung

In der LEMoN-Pipeline muss vor der eigentlichen Arbeit mit lokalen Grammatiken das jeweilige Arbeitskorpus vorbereitet werden. Dieser Prozess wird im Folgenden beschrieben. Die Schritte gelten für alle Arten von Korpora: Testkorpora, Validierungskorpora und für das Korpus des Hauptlaufs, vgl. Abschnitt 7.2.

Zunächst wird das Zielkorpus als Textdatei in Unitex importiert. Über die integrierte Vorverarbeitung wird der Text standardmäßig in Sätze und Token zerlegt. Die von Unitex mitgelieferte Satzenderkennung ist robust, bietet jedoch Raum für Optimierung. Für LEMoN kommt daher die elaborierte Satzenderkennung von Cibrán Beiras Cunqueiro zum Einsatz ([C. Beiras Cunqueiro 2005]). Nach dem Preprocessing steht das Korpus in Unitex zur weiteren Verarbeitung bereit.

8.2 Lexikographierung von unbekannten Wörtern

Unitex erzeugt nach der Vorverarbeitung eine Liste `ERR` aller Wörter, die in den als Default-Wörterbüchern der ausgewählten Sprache nicht verzeichnet sind, sogenannte OOV. Ziel ist es, den Anteil unbekannter Wörter im Korpus so weit zu senken, dass die darauf angewendeten Grammatiken stabil arbeiten. Die Prozedur ist halbautomatisch und stufig, mit Fokus auf die Wortarten `N`, `A` und `DET`. Andere Wortarten wurden nicht oder nur pragmatisch ausgezeichnet.

Ablaufplan:

1) Eingang: `ERR`-Liste pro Korpus kopieren und versionieren.

2) Vorabprüfung: Artefakte und Dubletten sichten, Reihenfolge festlegen.

3) Reihenfolge wählen

    i) Ausnahme zuerst: zunächst Ausnahmen markieren, damit anschließende breite Regeln zuverlässig greifen.

    ii) Breite Regel zuerst: erst breite Wortklassen klassifizieren, danach feinere Unterscheidungen vornehmen.

4) Operative Schritte

    i) Vorfilter Gremlins

    ii) Fremdsprachenfilter

iii) Affixbasierte Vorauszeichnung

iv) Orthographievarianten

v) Umgang mit Fremdwörtern

vi) Export

vii) Qualitätssicherung

Vorfilter Gremlins: Mit Gremlins bezeichnet man kleine, schwer zu findende Fehler oder Probleme in der elektronischen Textverarbeitung, beispielsweise ist der Begriff häufig in Verwendung für Zeichenkodierungsprobleme bei der Entwicklung von Websites. Auch bei der Lexikonauszeichnung finden sich Gremlins, da die verwendeten Korpora oft aus dem Web stammen und entsprechende Kodierungsfehler mit aufgenommen haben, oder durch Digitalisierung Artefakte entstanden sind. In Tabelle 20 (Anhang A) ist eine nicht abschließende Liste von Zeichen abgebildet, die Indikatoren dafür sind, dass die entsprechenden Zeilen aus dem Wörterbuch gelöscht werden können. Dadurch reduziert sich der Umfang gleich zu Beginn unter Umständen erheblich, was Lade- und Kompilationszeiten positiv beeinflusst und die Textmenge, die manuell kontrolliert werden muss, reduziert. Grenzfälle werden händisch geprüft und auffällige Einträge werden zusätzlich protokolliert, um spätere Nachvollziehbarkeit zu gewährleisten.

Fremdsprachenfilter: Nicht-spanische Einträge werden entfernt, ohne relevante spanische Lemmata zu verlieren. Bewährt haben sich folgende Schritte:

- Zunächst alle Wörter entfernen, die aus nicht-lateinischen Zeichen bestehen.
- Anschließend alle Wörter entfernen, die diakritische Zeichen enthalten, die im Spanischen nicht vorkommen (siehe Tabelle 22, Anhang A).
- Abschließend alle Wörter entfernen, deren unmittelbarer Kontext aus in Tabelle 21 (Anhang A) aufgeführten nicht-spanischen Funktionswörtern besteht.
- Überprüfung aller Wörter, die weitere Zeichen enthalten, die im Spanischen ungewöhnlich sind. Beispielsweise sind Bindestrich-Wörter im Deutschen, im Französischen und im Englischen weitaus häufiger als im Spanischen anzutreffen.

Affixbasierte Vorauszeichnung: Das Hauptziel besteht darin, eine möglichst große Menge von Wörtern effizient und korrekt mit einer Wortart zu taggen. Um dies zu erreichen, haben sich verschiedene Heuristiken bewährt, die systematisch und basierend auf dem jeweiligen Datensatz abgearbeitet werden sollten.

Zunächst können reguläre Ausdrücke verwendet werden, um die Wortliste schrittweise zu bearbeiten. Man beginnt mit Abfragen (Queries), die typischerweise nur Wörter der gewünschten Wortart erfassen, und schränkt dann sukzessive die verbliebenen Wörter weiter ein. Dabei sind Suffixe und - mit Einschränkungen - auch Präfixe von zentraler Bedeutung. Die Herangehensweise an Präfixe und Suffixe unterscheidet sich erheblich. Bei Präfixen ist in der Regel keine bestimmte Wortart vorgegeben, was bedeutet, dass ein Präfix nur unter bestimmten Umständen Hinweise auf die Wortart geben kann. Beispielsweise sind Wörter mit dem Präfix *super-* häufig Nomen (wie *Supermercado* oder *Superhéroe*), können jedoch auch Adverbien (*superbien*), Adjektive (*superior*) oder Verben (*superar*) sein.

Im Gegensatz dazu bieten einige Suffixe zuverlässigere Indikatoren für bestimmte Wortarten. Ein Beispiel dafür ist die Endung *-a*, die häufig für weibliche Singularformen verwendet wird. Allerdings ist die Fehlerquote in diesem Fall aufgrund der Vielzahl an Nicht-Nomen, die ebenfalls auf *-a* enden, sehr hoch. Eine Verfeinerung könnte die Endung *-ita* sein, die voraussichtlich weibliche Nomen im Diminutiv bezeichnet. Dennoch gibt es auch hier noch Wörter, die auf *-ita* enden, aber keine Diminutive oder Nomen sind, wie Adjektive (z. B. *favorita*) oder konjugierte Verben (*visita*).

Eine noch präzisere Submenge wäre die Endung *-cita*, da diese seltener vorkommt und somit Fehler leichter nachvollzogen und korrigiert werden können. In jedem Fall ist eine manuelle Kontrolle unerlässlich, um die Korrektheit der Zuordnungen sicherzustellen. Durch die vorherige Überprüfung der Endungen auf *-ita* könnte die Gesamtzahl der verbliebenen Wörter bereits signifikant reduziert werden. Auf diese Weise arbeitet man sich von Suffix zu Suffix vor, bis nur noch große Klassen von Wörtern übrig bleiben. Diese Klassen sind jedoch durch die vorherige Selektion um viele eindeutige Fälle reduziert worden, sodass sie handlicher sind und nun leichter manuell getaggt werden können, siehe Tabellen 23 und 24 sowie Gegenbeispiele in Tabellen 25 und 26 (jeweils in Anhang A). Zusätzlich kann bei Bedarf eine stichprobenartige Gegenprüfung der bereits getaggten Teilmengen erfolgen, damit Fehlklassifikationen früh erkannt und zurückgerollt werden können.

Orthographievarianten: Im Unterschied dazu kann man argumentieren, dass spanische Wörter, die aufgrund eines Rechtschreibfehlers (Auslassung, Einfügung, Verdrehung, falscher oder fehlender Akzent) nicht der Standardorthographie entsprechen, als Varianten des standardsprachlichen Lemmas ausgezeichnet werden. Die Entscheidung darüber hängt mit Sicherheit auch mit dem jeweiligen Fehler zusammen. Entscheidende Faktoren können etwa die Distanz zum Ausgangswort oder die Häufigkeit des Fehlers sein. Gerade im Bereich nachträglich digitalisierter Medien trifft man beispielsweise häufig auf ähnliche, dem Digitalisierungsprozess geschuldete Fehler. Für den praktischen Ablauf hat es sich bewährt, wiederkehrende Fehlerbilder getrennt zu sammeln, damit sie konsistent behandelt werden können.

Umgang mit Fremdwörtern und Fachausdrücken: Fremdsprachliche (Fach-)Ausdrücke können bei Bedarf lexikographiert werden, da sie insbesondere in Fachsprachen einen erheblichen Teil spanischer MWEs ausmachen - in der Alltagssprache hingegen wenig bis gar nicht. Hier werden sie daher nur aufgenommen, wenn es sich um stark eingebürgerte Formen handelt.

Export: Verbleibende Einträge werden entweder minimal lexikographiert (alternativ können sie unausgezeichnet in ERR bleiben und nicht Teil des Wörterbuchs werden). Ausgezeichnet wird mindestens ein POS-Tag. Bei Nomen lassen sich oft Numerus und Genus einfach automatisiert auszeichnen.

Qualitätssicherung: Sowohl während der Arbeit als auch am Ende sollten die ausgezeichneten Wörter händisch evaluiert werden, um eine korrekte Auszeichnung zu gewährleisten. Es empfiehlt sich, die Prüfschritte zu protokollieren und Beispiele für Grenzfälle festzuhalten, damit spätere Entscheidungen auf konsistenter Grundlage getroffen werden können.

Aus pragmatischen Gründen wurde für diese Arbeit auch eine den Regeln der Nachhaltigkeit widersprechende Konvention getroffen: Verben, Adjektive, Adverbien und Präpositionen sollen zwar möglichst vollständig klassifiziert werden. Selbiges trifft auf Nomen zu, die sich separat aufgrund ihrer Endung finden und mit syntaktischen Informationen wie Numerus und Genus auszeichnen lassen.

Als zeitweilige Fallback-Policy wurden jedoch übrige unbekannte Einheiten der Einfachheit halber als Nomen ausgezeichnet. Diese Entscheidung birgt ein Risiko für die Qualität einzelner Matches und wurde daher nicht durchgehend angewendet.

Für den hier verfolgten Zweck ist sie jedoch vertretbar, weil die zu erwartenden Verzerrungen angesichts der Datenmenge und der relativen Vorkommenshäufigkeiten gering ins Gewicht fallen.

Diese pragmatische Herangehensweise kann damit gerechtfertigt werden, dass es nicht das Ziel dieser Arbeit ist, perfekte Arbeitslexika anzulegen, sondern ein stabiles und bewältigbares Verfahren zur Extraktion von MWEs aus Korpora sowie ein entsprechendes MWE-Lexikon zu entwickeln.

Einbinden als neues Wörterbuch: Nachdem die Auszeichnung und Überprüfung der unbekannten Wörter des Testkorpus abgeschlossen ist, kann das neu entstandene Wörterbuch in Unitex eingebunden und auf das Korpus angewendet werden. Die Anzahl der unbekannten Wörter hat sich entsprechend reduziert. Gegebenenfalls bleiben fremdsprachliche oder fehlerhafte Wörter weiterhin als unbekannt übrig. In diesen Fällen wird die OOV-Liste aktualisiert und die betreffenden Einträge werden in einer weiteren Iteration gezielt nachbearbeitet.

8.3 Testphase

Die Testphase beginnt mit einem vorverarbeiteten Testkorpus samt OOV-Wörterbuch. Die Arbeit mit lokalen Grammatiken stellt den Hauptteil dieses Arbeitsschritts der LEMoN-Pipeline dar. In diesem Schritt werden lokale Grammatiken entwickelt, die das Phänomen der MWEs so modellieren, dass diese aus verschiedenen Textkorpora extrahiert und lexikographiert werden können. Die Erstellung lokaler Grammatiken ist ein komplexer Prozess, der auf der Analyse und Modellierung linguistischer Phänomene beruht.

Der erste Schritt dieser Arbeit ist eine intuitive, auf linguistischen Beobachtungen und Vorüberlegungen basierende Annäherung an das zu modellierende Phänomen in Form einer einfachen lokalen Grammatik. Diese wird dann in Unitex auf das Testkorpus angewendet. Unitex gibt eine Konkordanz zurück, die die über die lokale Grammatik identifizierten Stellen samt linkem und rechtem Kontext darstellt.

Im nächsten Schritt werden die gefundenen Elemente mit der gewünschten Erkennung verglichen: Werden mehr Elemente erkannt, als gewünscht? Werden andere Elemente nicht erkannt, obwohl sie es sollten? Diese Beobachtungen werden in die lokale Grammatik eingearbeitet. Dieser Schritt ist notwendig, um die getroffenen Annahmen über die Ausprägung eines linguistischen Phänomens mit einem anderen Teil der sprachlichen Wirklichkeit gegenprüfen zu können. Erkenntnisse aus dieser

Arbeit können sein:

Bestätigung: Das Phänomen ist zutreffend modelliert. Es finden sich keine oder wenige Ausnahmen, die eine Anpassung der Modellierung erforderlich machen.

Übergeneralisierung: Das Phänomen ist zu stark generalisiert worden. Es finden sich viele unzutreffende Ergebnisse in der Konkordanz wieder. Die lokalen Grammatiken müssen an dieser Stelle so ummodelliert werden, dass das Phänomen präziser getroffen wird.

Überspezifisch: Das Modell des Phänomens ist zu stark eingeschränkt. Es finden sich zahlreiche Vorkommen im neuen Text, die nicht von der lokalen Grammatik erfasst werden. Entsprechend müssen die Graphen umgearbeitet werden, um auch die neuentdeckten Fälle abzudecken (vgl. Abschnitt zu Bootstrapping 5.3).

Anhand der gefundenen Ergebnisse und mithilfe der linguistischen Intuition des Bearbeiters wird der Graph verfeinert. Die Ausgabe des modifizierten Graphen zeigt die herbeigeführten Veränderungen und neu gefundenen Ergebnisse in Form einer Konkordanz an und legt dabei Möglichkeiten der Verbesserung offen, die der Anwender im nächsten Arbeitsschritt berücksichtigen kann.

Letztlich gibt es in der Arbeit mit lokalen Grammatiken wohl selten den Moment, in dem ein Graph im abschließenden Sinne vollständig ist. Auf diese Weise lassen sich jedoch Zustände erreichen, in denen weitere Arbeiten am Graphen mehr neue Probleme schaffen als sie lösen. Spätestens in diesem Moment ist es ratsam, entweder ein anderes Testkorpus zu verwenden, ein Validierungskorpus heranzuziehen oder eine lokale Grammatik als ausreichend zu betrachten und die Arbeit an ihr einzustellen.

Da die Analyse von MWEs eine Aufgabe ist, die die Konstruktion einer großen Anzahl von Graphen erfordert, ist neben der Entdeckungsmethode der lokalen Grammatiken auch eine übergeordnete Organisation der Graphen hilfreich, um beste Ergebnisse zu erzielen. Die Qualität der gefundenen Ergebnisse und damit des Lexikons steht und fällt mit der Qualität der zugrunde liegenden lokalen Grammatiken. Je nach Herangehensweise bieten sich Organisationsformen etwa nach semantischen, syntaktischen oder pragmatischen Kategorien an. Die für diese Arbeit verwendete Organisation der Graphen wird in Kapitel 9 beschrieben.

Nachdem die Qualitätskontrolle zu dem Ergebnis kommt, dass keine substanziellen Verbesserungen mehr herbeigeführt werden können, endet die Testphase mit dem ersten Testkorpus. Es können mehrere verschiedene kleinere Korpora verwendet werden, bis ein allgemeines Gefühl für die Qualität der lokalen Grammatiken entwickelt ist und dieses gegen ein größeres, gegebenenfalls domänenfremdes Validierungskorpus getestet werden kann.

8.4 Validierungsphase

Zur Validierung und Härtung der lokalen Grammatiken werden diese auf das Validierungskorpus angewendet. In der schematischen Darstellung der LEMoN-Pipeline auf Abbildung 8 entspricht dies der Validierungsphase. Die Validierungskonkordanz wird mit demselben Blick betrachtet wie die Testkonkordanzen im Bootstrapping. Das größere und oft domänenfremde Korpus stellt sicher, dass die lokalen Grammatiken auf eine andere Datengrundlage treffen und die zuvor getroffenen Annahmen mit einem weiteren Teil der sprachlichen Wirklichkeit abgeglichen werden.

Wenn die Validierungskonkordanz akzeptable Resultate in der Qualitätsprüfung zeigt, fällt diese positiv aus und die Grammatiken werden als gehärtete lokale Grammatiken in die nächste Phase übernommen. Fällt sie negativ aus, werden die Grammatiken verfeinert und erneut auf das Validierungskorpus angewendet. Dies kann eine einmalige Anpassung sein, es kann sich aber auch um eine zweite Bootstrapping-Phase handeln, vgl. den Rücksprungpfeil in Abbildung 8. An dieser Stelle kann auch das Validierungskorpus gewechselt werden, damit die Verallgemeinerbarkeit der Regeln weiter verbessert wird.

Der Validierungszyklus kann wiederholt werden, bis die Graphen hinreichend gehärtet erscheinen. Ein praktikabler Endpunkt ist erreicht, wenn wiederholte Modifikationen keine substanziellen Verbesserungen mehr bringen und neue Anpassungen eher neue Probleme erzeugen. In diesem Stadium wird der Validierungslauf gegebenenfalls noch einmal durchgeführt und dann abgeschlossen.

Wie beschrieben gibt es keinen fixen Punkt, an dem keine Verbesserungen mehr denkbar sind. Es ist jedoch wichtig, die lokalen Grammatiken nicht immer weiter mit feineren Verästelungen zu versehen, damit Verbesserungen auf der einen Seite nicht durch Seiteneffekte insgesamt zur Verschlechterung führen oder die Grammatik nicht mehr lesbar ist.

Gelangt die Qualitätskontrolle zu dem Ergebnis, dass mit keiner weiteren substanziellen Verbesserung zu rechnen ist, kommt die Validierung zum Ende und die

lokalen Grammatiken sind bereit für den Hauptlauf.

8.5 Hauptlauf und Export

Wenn die Arbeit mit den lokalen Grammatiken abgeschlossen ist, ist es an der Zeit, sich den ungleich umfangreicheren Trainingsdaten zuzuwenden: LEMoN wird auf dem EuroParl-Korpus (vgl. Abschnitt 7.2) erstellt. Dieses ist aufgrund seiner Größe auf mehrere Dateien aufgeteilt. Im Unterschied zum bisherigen Verfahren werden alle zuvor manuell in Unitex ausgeführten Operationen nun über die Kommandozeile oder skriptbasiert auf allen Dateien automatisch ausgeführt. Zum Einsatz kommen die in der Validierung gehärteten lokalen Grammatiken.

Zunächst werden die Korpora wie oben beschrieben aufbereitet. Im Anschluss wird ein Skript gestartet, das für alle Dateien mit den gewählten Einstellungen die Programme `Locate` und `Concord` in Unitex ausführt und die Ergebnisse - das heißt die Konkordanzen `concord.txt` - speichert.

Aufgrund der Aufteilung der Trainingsdaten auf mehrere Dateien liegen die Konkordanzen und die daraus gewonnenen Kandidatenlisten zunächst verteilt vor und werden skriptbasiert zusammengeführt. Durch die entsprechend angelegten Transducer-Outputs ist die zusammengeführte Datei automatisch eine gültige DELA-Datei. Dabei werden Dubletten zunächst nicht überschrieben, um statistische Daten über die Häufigkeit auftretender Kandidaten oder POS-Muster nicht zu verzerren. Die zusammengeführte Liste bildet das MWE-Kandidateninventar der Pipeline.

Im nächsten Schritt wird ein Skript zur Deduplikation ausgeführt, das jede Oberflächenform mit einem einmaligen POS-Muster auf ein einzelnes Vorkommen reduziert. Im selben Schritt werden für die Auswertung statistische Daten erhoben (siehe Kapitel 10). Das Ergebnis wird als `LEMoN.dic` gespeichert und steht damit für die weitere Verarbeitung zur Verfügung.

Als letzter Schritt, der außerhalb des Fokus dieser Arbeit liegt, muss die Lexikographierung der Kandidaten in `LEMoN.dic` erfolgen: Für jeden einzelnen Kandidaten muss geprüft werden, ob es sich um eine MWE handelt und ob die provisorisch zugewiesenen POS-Tags übernommen werden können oder angepasst werden müssen. Kandidaten mit hoher Frequenz und ohne alternative Analysen sind ein geeigneter Startpunkt. Kandidaten mit vielen alternativen Lesarten oder geringer Frequenz erfordern zusätzliche Recherchezeit.

Mit dem erfolgreichen Abschluss dieser Schritte ist die erste Version des Lexikons

`LEMoN.dic` entstanden. Durch das systematische Lexikographieren mithilfe der heuristischen Ansätze ist eine gleichbleibende hohe Qualität gewährleistet. Der Einsatz der lokalen Grammatiken stellt sicher, dass in der Regel ein großer Teil der in den Trainingskorpora belegten spanischen nominalen MWEs identifiziert und lexikalisiert wurde. Dies bietet eine solide Basis für zukünftige Erweiterungen, und weitere Studien mit noch größeren Korpora oder in anderen sprachlichen Kontexten können mit dem Lexikon durchgeführt werden.

Im nächsten Kapitel werden die LEMoN-Grammatiken im Detail besprochen. Die Evaluation von `LEMoN.dic` und der Pipeline findet im daran anschließenden Kapitel statt.

9 LEMoN-Grammatiken

The simplest is always the most difficult.

Masaaki Hatsumi

Im vorigen Kapitel wurde beschrieben, wie die LEMoN-Pipeline aufgebaut und ausgeführt wird, vgl. insbesondere Abbildung 8. Ein integraler Bestandteil dieses Prozesses ist der Einsatz lokaler Grammatiken, mit denen aus Korpora MWEs extrahiert werden.

Die Theorie und der Entwicklungsprozess lokaler Grammatiken werden in Kapitel 5 dargestellt. Dieses Kapitel beschreibt dagegen Aufbau, Organisation und Analyse der im Rahmen dieser Arbeit entwickelten LEMoN-Grammatiken.

9.1 Hinweise zum Lesen lokaler Grammatiken

Es ist leicht, lokale Grammatiken zu entwickeln. Es ist hingegen sehr herausfordernd, *gute* lokale Grammatiken zu entwickeln. Das Lesen lokaler Grammatiken ist ebenfalls leicht - die folgenden Anmerkungen erläutern, wie die einzelnen Graphen und Knoten in den folgenden Abschnitten zu interpretieren sind.

- Lokale Grammatiken werden von links nach rechts gelesen. Der Text läuft vom Startknoten über Zwischenknoten zum Endknoten.
- Als erkannt gilt nur Text, der den Endknoten erreicht. Erkannte Einheiten können als Konkordanz mit linkem oder rechtem Kontext ausgegeben oder alternativ in eine Datei geschrieben werden.
- Zwischenknoten können Portale in Subgraphen sein. Subgraphen besitzen jeweils eigene Start- und Endknoten und können weitere Subgraphen enthalten. Nach dem Endknoten eines Subgraphen wird in den Obergraphen zurückgeführt.
- Knoten dürfen Schleifen auf sich selbst besitzen und so Rekursion erzeugen. Größere verschachtelte Strukturen heißen Graphenbäume.
- Berücksichtigt werden nur Pfade, die vom Startknoten aus erreichbar sind und zu einem Endknoten führen. Ganze Graphenbäume lassen sich durch Entfernen einer Verbindung gezielt aus der Analyse nehmen, zum Beispiel zu Testzwecken oder zum Ein- und Ausblenden bestimmter Strukturen.

- Die Knoten zwischen den gerichteten Kanten können einen einzigen Pfad bilden oder sich verzweigen, auch mit Schleifen.
- Zwischenknoten können Annotationen tragen. Sie erscheinen rot im Editor und können auf Wunsch in die Ausgabe übernommen werden. In diesem Fall dienen sie als Transducer-Output und werden zusätzlich zum Treffertext oder ersetzend ausgegeben, sobald die markierte Transition passiert wird.
- Zwischenknoten können Filter enthalten. Die in den LEMoN-Grammatiken verwendeten Filtertypen sind in Tabelle 9 dargestellt.

Tabelle 9: Filtertypen in den LEMoN-Grammatiken und ihre Funktion

Filtertyp	**Beschreibung**
`<E>`	Epsilon- bzw. Leertransition. Der Übergang erfolgt ohne Tokenabgleich und modelliert die Abwesenheit eines Konnektors. Optionalität wird als Alternative zwischen `<E>` und einem expliziten Konnektor realisiert.
Literal	Feste Oberflächenform im Übergang. Das Matching ist tokenbasiert und setzt Wortgrenzen voraus.
Subgraph	Aufruf eines Untergraphen, z. B. `:syn_all` oder `..:auxi:prep`. `:` referenziert einen Unterordner, `..:` einen übergeordneten Ordner.
`<N>`, `<A>`, `<DET>` …	POS-Klassen aus den aktivierten Lexika. Auch komplexe Einheiten werden akzeptiert, wenn sie im Lexikon entsprechend gekennzeichnet sind.
`<N:fs>`, `<A:mp>` …	POS-Klassen mit morphologischen Merkmalen, z. B. `fs` = feminin Singular und `mp` = maskulin Plural, sofern im Lexikon kodiert.
Alternativlisten	Mehrere zulässige Optionen in einem Knoten. Im Editor als einzelne Zeilen notiert und logisch als Alternativen interpretiert. Erlaubt sind u. a. Literale, POS-Klassen und `<E>`. Optionalität entsteht durch Aufnahme von `<E>` oder durch einen alternativen Graphenpfad. Beispiel: `de | del | de la | <DET> | <E>`.

In LEMoN nutze ich diese Möglichkeit, um bereits im Graphenlauf die DELA-Syntax zu erzeugen. Erfasst werden die Oberflächenform und die erkannten POS-Muster, etwa `NN` oder `NANN`. Der Export liefert damit fast ohne weitere Nachbearbeitung DELA-kompatible Kandidatenzeilen.

9.2 Übersicht

Leitidee: Die LEMoN-Grammatiken identifizieren spanische nominale MWEs, indem sie wiederkehrende syntaktische Muster als lokale Grammatiken modellieren und daraus Kandidaten extrahieren. Der Ansatz ist recall-orientiert und arbeitet mit transducerbasierten Annotationen für den direkten DELA-Export.

> [...] it appears that most of the terms encountered are noun phrases corresponding to a limited number of syntactic patterns. These patterns define the morphosyntactic structures of multi-words units which are likely to be terms. We define the length of a MWU as the number of main items it contains. A main item is either a noun, an adjective, a verb or an adverb. Thus, neither determiners nor prepositions are considered main items.
>
> [B. Daille 1994, 515]

Terminologie: In dieser Arbeit bezeichnet POS-Muster das reduzierte Muster über den MIs eines Kandidaten. Als MIs gelten ausschließlich Nomen `N` und Adjektive `A`. Funktionselemente wie Präpositionen `PREP` und Determinierer `DET` werden nicht gezählt. Beispiel: *compañero de piso* hat die Oberflächenfolge `N PREP N`. Das POS-Muster ist `NN` mit 2 MIs. Wenn nicht das MI-basierte POS-Muster gemeint ist, sondern die vollständige Abfolge aller POS-Tags, referenziere ich darauf als *vollständige POS-Folge*. Jeder Eintrag in `LEMoN.dic` endet mit `MI=` und dem entsprechenden (reduzierten) POS-Muster.

Erkennungsziel und Abgrenzungen: LEMoN zielt auf spanische nominale MWEs mit bis zu 5 MIs. Erfasst werden `<N>` und `<A>` sowie entsprechend ausgezeichnete komplexe Einheiten. Ambige Tokens, die in den Lexika auch als `<N>` oder `<A>` lesbar sind, werden zunächst zugelassen. Wo möglich, reduziere ich solche Fälle über Kontextbedingungen. Verbleibende Ambiguitäten müssen in der nachgelagerten Qualitätskontrolle behandelt werden.

9.3 Architektur

Lauf auf hoher Ebene:

1) **Start):** `start.grf` leitet unmittelbar an die syntaktische Steuerung weiter.

2) **Syntaktische Steuerung:** `syn_all` verzweigt in `syn_typ` als Basispfad und in `syn_atyp`, das im Hauptlauf deaktiviert und in der Ablation getestet ist. Der Einstieg für `syn_typ` prüft zwei Fälle. Entweder ist der Kandidat durch korrespondierende Begrenzungszeichen eingerahmt (z. B. „...“, (...) , ¿...? , ¡...!), oder es steht ein Determinierer davor (`det_all.grf`).

3) **MI-Ebene:** verzweigt in die Graphen für 2, 3, 4 oder 5 MI, in denen die jeweiligen Kandidatenpfade aufgerufen werden.

4) **Erkennungsebene:** Je POS-Muster ist ein Graph modelliert, zwischen den einzelnen MIs sind Konnektoren und Präpositionalkonstruktionen über Hilfsgraphen eingebunden. Jeder Graph ist mit Transducer-Output annotiert.

5) **Hilfsgraphen (auxi):** Es gibt zwei Familien: Die `DET`-Hilfsgraphen (`det_complex`, `det_composed`, `det_numeral`, `det_prozent`, `num_all`) kommen im `det_all.grf` zum Einsatz. Die Konnektor-Hilfsgraphen (`btwNN`, `btwNA`, `btwAN`, `btwAA`, `prep`, `de_var`, `a_var`, `cnx`) stehen innerhalb der Kandidatenpfade.

6) **Annotation und Export:** Der Transducer schreibt Oberflächenform und reduziertes MI-Muster in DELA-kompatible Zeilen, etwa `MI=NN`.

Ebenen des Laufs (Schichtenmodell): Abbildung 9 fasst diese fünf Ebenen zusammen.

Klassen der LEMoN-Grammatiken:

- Laufgraphen (Steuerung)
  - `start.grf`
  - `syn_all.grf`
  - `syn_{2,3,4,5}_Main_Items.grf`
- Erkennungsgraphen (Kandidaten)
  - konkrete Muster wie `NN`, `NA`, `NNA` ... mit Transducer-Ausgabe
- DET-Hilfsgraphen
  - `det_complex`, `det_composed`, `det_numeral`, `det_prozent`, `num_all`

- Konnektor-Hilfsgraphen
  - `btwNN`, `btwNA`, `btwAN`, `btwAA`, `prep`, `de_var`, `a_var`, `cnx`
- Export/Transducer
  - Ausgabe der Form und `MI=...` im DELA-Format

Für eine schematische Darstellung siehe Abbildung 9. Durchgezogene Pfeile zeigen den Hauptfluss der LEMoN-Grammatiken. Gestrichelte Pfeile markieren optionale Hilfsgraphen. Der Eintritt erfolgt über den Startgraphen, entweder durch ein Begrenzungszeichen-Paar oder durch einen Determinierer (`det_all`).

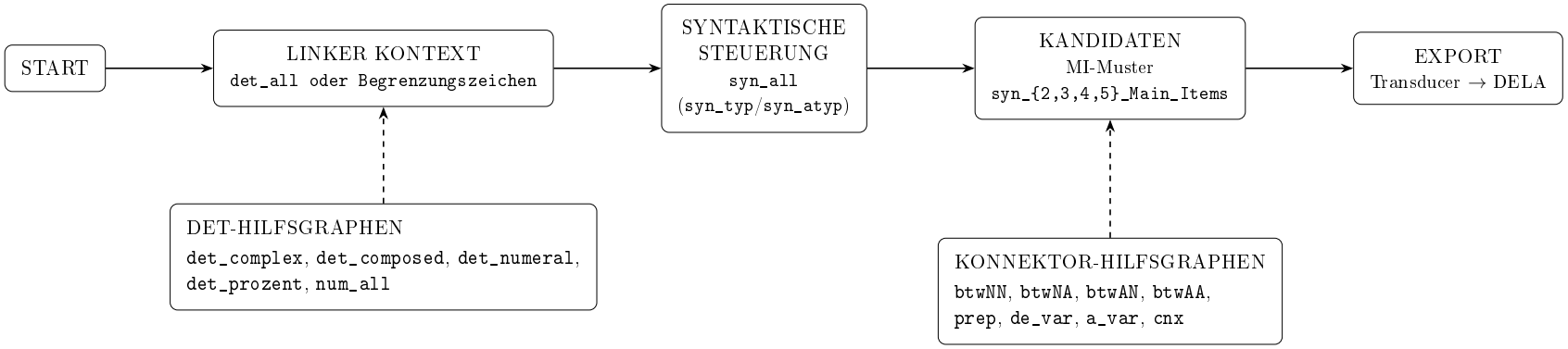

Abbildung 9: Schematische Darstellung der LEMoN-Grammatiken

9.4 Detaillierte Besprechung

9.4.1 Start

Die Anwendung der lokalen Grammatiken beginnt mit dem Startgraphen `start.grf`. Er bietet die Option, einen linken oder rechten Filter für die syntaktische Erkennung zu definieren, für diese Arbeit ist er aber so konfiguriert, dass er direkt zu der als Subgraph eingefügten syntaktischen Erkennung `syn_all.grf` weiterleitet.

9.4.2 Syntaktische Steuerung

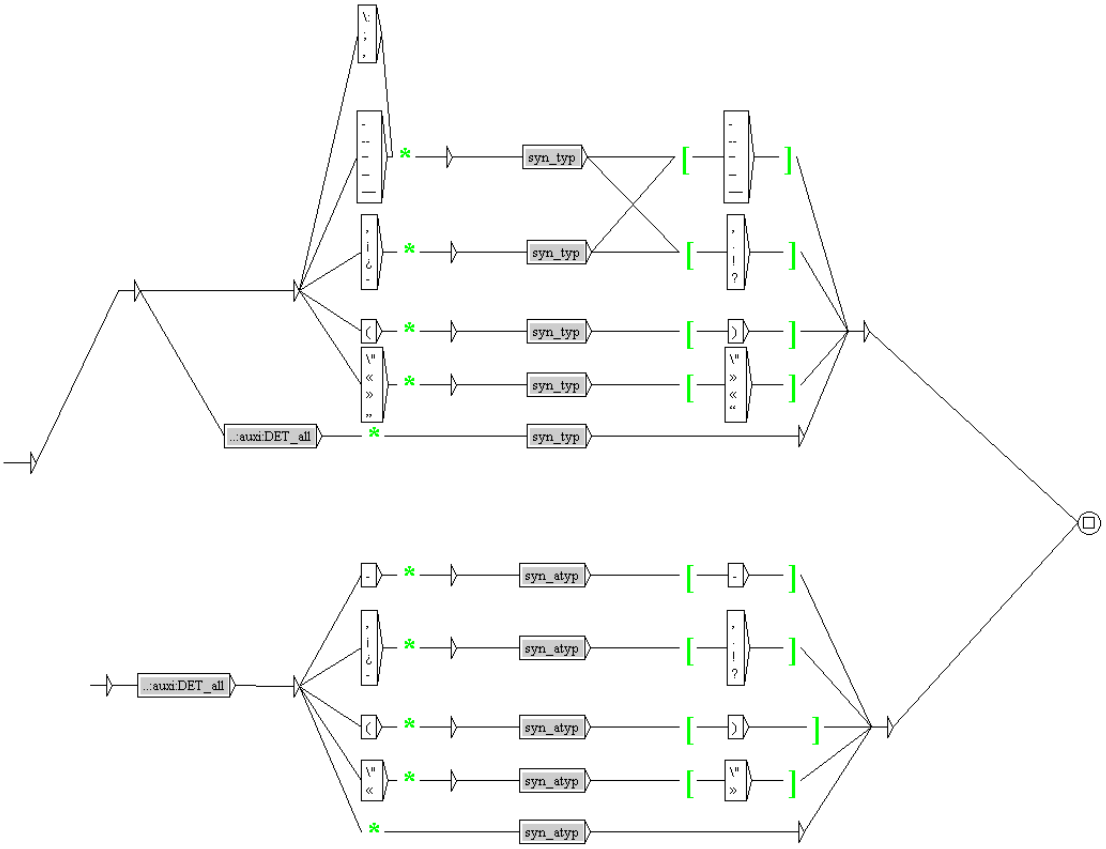

Abbildung 10: `syn_all.grf`

Der Graph in Abbildung 10 besteht aus zwei Zweigen: typischen und atypischen syntaktischen Strukturen. Der Zweig für atypische Strukturen ist standardmäßig deaktiviert und wird nur in der Ablationsstudie zugeschaltet. Im typischen Bereich gibt es zwei alternative Einstiege. Erstens werden Kandidaten zugelassen, wenn sie durch

ein Paar korrespondierende Begrenzungszeichen eingerahmt sind, etwa durch Anführungszeichen („. . .“ bzw. «. . . »), Klammern oder ein Gedankenstrichpaar. Zweitens, falls kein Begrenzungszeichenpaar vorliegt, dient ein vorangestellter Determinierer als Filter, im Subgraphen `det_all` definiert.

Für das Spanische sind die umgedrehten Ausrufe- und Fragezeichen besonders nützlich, zum Beispiel *¿Gas natural?* und *¡Cielo santo!*.

Im Editor markieren die grünen Klammern und Sternchen den Kontext: Alles links des grünen Sternchens sowie innerhalb der grünen eckigen Klammern wird nicht mit ausgegeben.

9.4.3 Determinierer

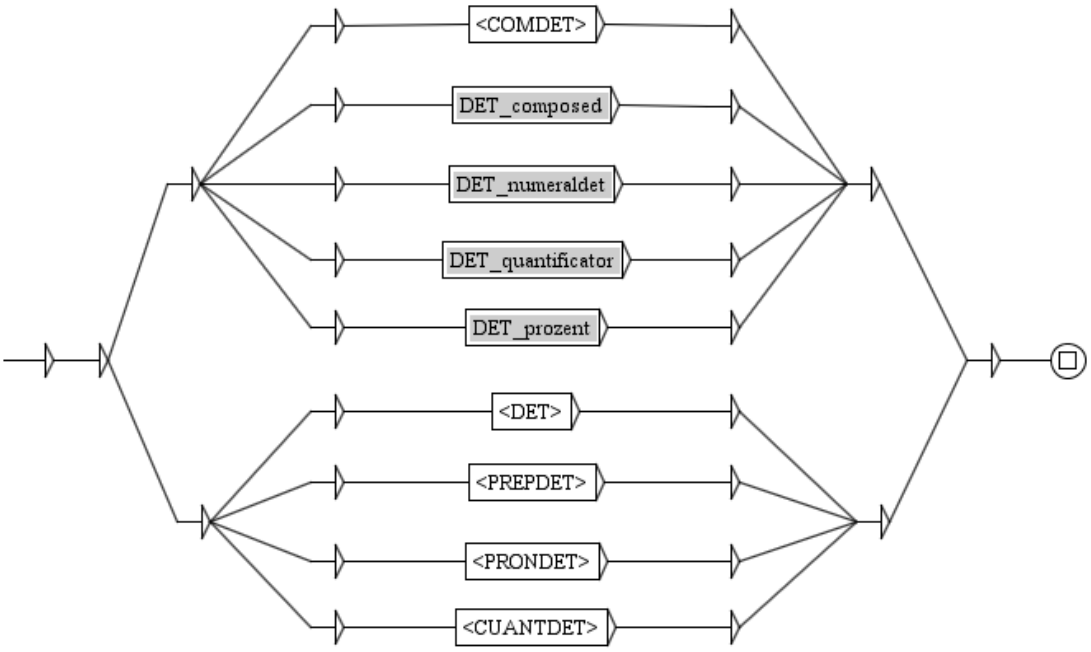

Abbildung 11: `det_all.grf`

Der Graph in Abbildung 11 erkennt spanische Determinierer über zwei Wege: Lexikontags und lokale Grammatiken für produktive mehrteilige Muster. Abgedeckt sind

1) **Einfache Determinierer** `<DET>`
   Beispiele: *el*, *los*, *un*

2) **Zusammengezogene Artikel** `PREPDET`
Beispiele: *del*, *al*

3) **Pronominale Determinierer** `<PRONDET>`
Beispiele: *mi*, *mis*, *tu*, *tus*, *su*, *sus*

4) **Quantifizierende Determinierer** `DET_quantificator`, `CUANTDET`
Beispiele: *todas*, *varios*, *bastante*

5) **Numerale Determinierer** `DET_numeraldet`
Beispiele: *tres*, *diecinueve*, *millones de*

6) **Prozentangaben als Determinierer** `DET_prozent`
Beispiele: *un 30 por ciento de*

7) **Komplexe determinierende Ausdrücke** `COMDET`, `DET_composed`
Beispiele: *cualquiera de los*, *algunos de los*, *el gran número de*, *la mayor parte de*, *una especie de*

9.4.4 Typische syntaktische Strukturen

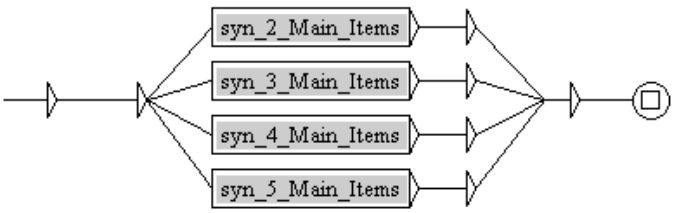

Abbildung 12: `syn_typ.grf`

Die typischen MWEs sind nach der Zahl der MIs organisiert. In dieser Arbeit werden Sequenzen mit mindestens 2 und höchstens 5 MIs berücksichtigt. Damit sind die meisten spanischen nominalen MWEs sowie ein großer Teil geschachtelter Fälle abgedeckt.

Die aufgerufenen Subgraphen sind:

- `syn_2_Main_Items`

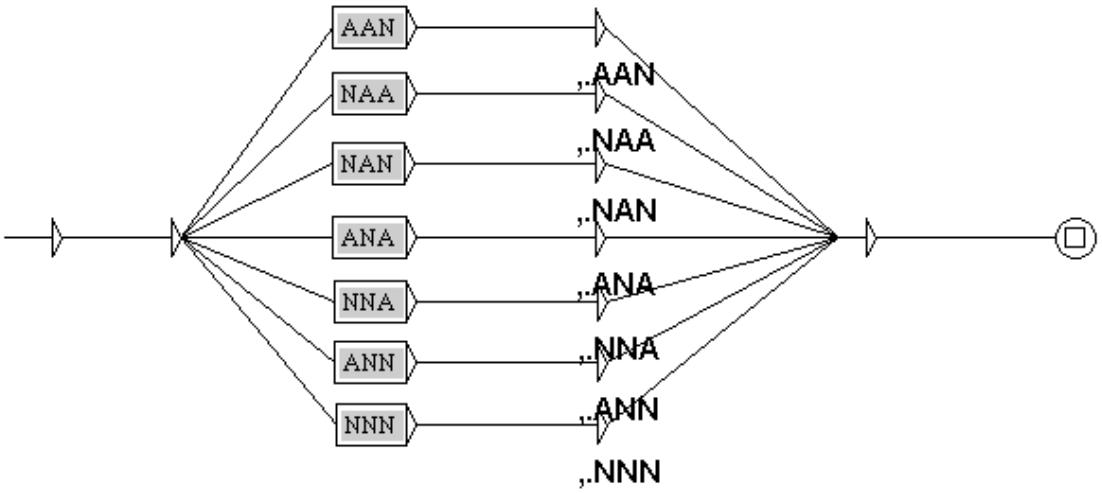

Abbildung 13: `syn_3_Main_Items.grf`

- `syn_3_Main_Items`
- `syn_4_Main_Items`
- `syn_5_Main_Items`

Diese Subgraphen decken die zulässigen Kombinationen aus Nomen `N` und Adjektiven `A` ab, mindestens ein `N` ist stets enthalten, das heißt, es sind auch extrem unwahrscheinliche Kombinationen mit meheren führenden `A` modelliert. In der Basis-Konfiguration ist der `AN`-Graph aus Entwicklungsgründen deaktiviert (vgl. Abschnitt 10.4.2). Abbildung 13 zeigt `syn_3_Main_Items.grf` exemplarisch.

Alle zulässigen Gruppierungen werden parallel erkannt und für den Export normalisiert. Eine automatische Zuordnung von Kopf und Basis erfolgt nicht und kann bei Bedarf nachgelagert kuratorisch vorgenommen werden.

Nominale MWEs erscheinen im Gebrauch häufig in Varianten, vor allem durch Einfügung, Permutation oder Koordination. Die LEMoN-Grammatiken decken diese Fälle über die MIs-Muster für 2 bis 5 MIs sowie über die eingebundenen Konnektoren wie `btwNN` und `btwNA` ab. Eine automatische Zuordnung von Kopf und Basis erfolgt nicht. Zur systematischen Einordnung vgl. [B. Daille 1996].

9.4.5 Atypische syntaktische Strukturen

Neben den regulären N/A-Kombinationen gibt es eine breite Klasse atypischer Muster, die eigene Grammatiken erfordert. Die Auswahl orientiert sich an der Taxonomie von [M. Mathieu-Colas 1996]. Ich habe daraus die Typen übernommen, die nicht schon durch die typischen Strukturen abgedeckt sind, und sie für das Spanische angepasst. Kategorien, die zwischen Nomen und Eigennamen unterscheiden, entfallen, weil die verwendeten Lexika beide als N kennzeichnen.

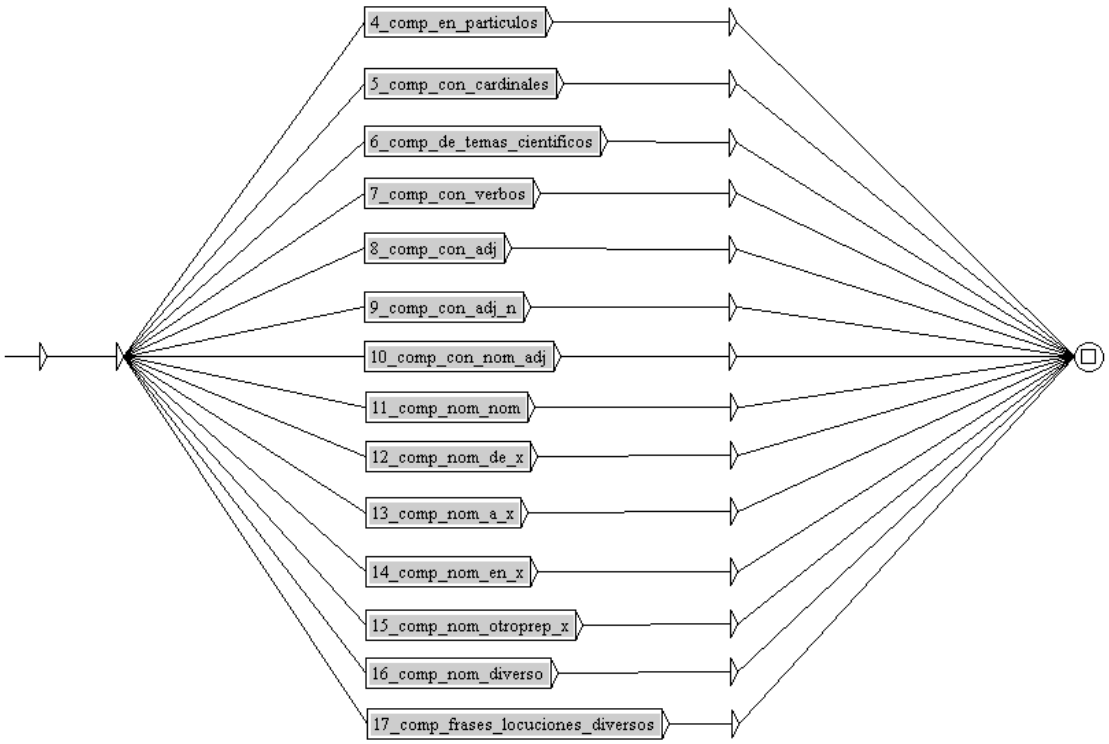

Abbildung 14: `syn_atyp.grf`

Die atypischen MWEs sind nach der Arbeit von [M. Mathieu-Colas 1996] folgend organisiert und nummeriert. Die ausgelassenen Nummern sind bereits über die typischen Strukturen abgedeckt.

Für LEMoN sind die folgenden Gruppen von atypischen syntaktischen MWE-Strukturen in insgesamt 14 Graphen als lokale Grammatiken modelliert:

1) **MWE mit Partikeln**
   Oberbegriff für oberflächlich unveränderliche Wörter wie Präpositionen oder Adverbien, etwa *a todo motor*.

2) **MWE mit Kardinalzahlen**
Ausdrücke mit Zahlangaben, etwa *quinto pino*.

3) **MWE mit Verben**
Verbhaltige Bildungen wie *tabla de planchar* oder *máquina de escribir*, siehe Abbildung 15.

4) **MWE mit Adjektiven**
Adjektivzentrierte Muster wie *blanco roto*.

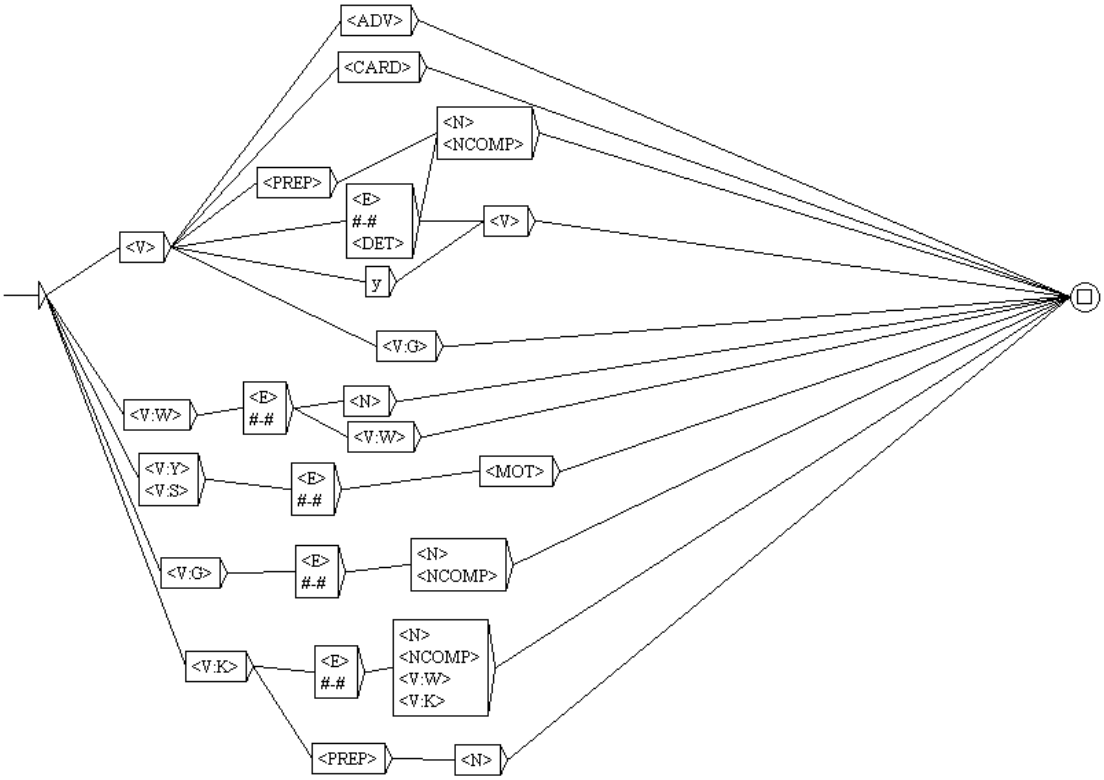

Abbildung 15: `compuestos-con-verbos.grf`

Präpositionale Varianten wie `N de N` fallen bei LEMoN unter die typischen syntaktischen Kombinationen und werden dort erfasst. Der atypische Zweig ist in der Basis-Konfiguration nicht an den Start angeschlossen und damit deaktiviert. Grund hierfür ist die geringere Präzision dieser Extraktion. Eine gesonderte Auswertung der atypischen Strukturen ist möglich, ohne die Ergebnisse der typischen Muster zu beeinflussen. In den Ablationsläufen habe ich die atypischen syntaktischen Gra-

phen getestet (und zwar in den Läufen L3, L3b, L5). Schon diese Teilaktivierung verschiebt das Längenprofil stark hin zu 5-MI, vgl. Abschnitt 10.5.

9.4.6 Kandidatengraphen

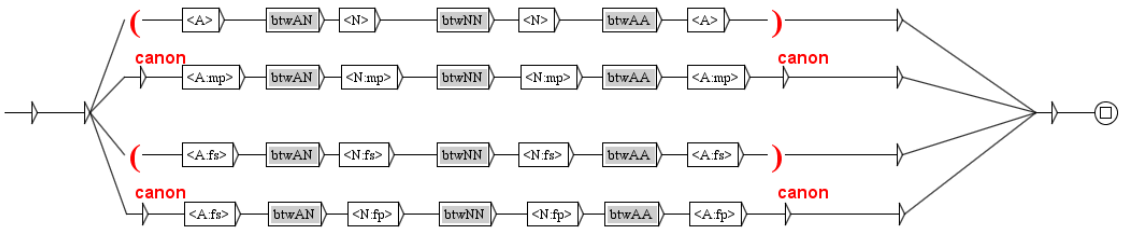

Abbildung 16: `ANNA.grf`

Die Kandidatengraphen sind nach dem jeweiligen Muster benannt, etwa `NN` oder `ANNA`. Für jeden Treffer erzeugen sie eine Transducer-Ausgabe. Die erlaubten Zwischenelemente werden über Konnektor-Hilfsgraphen (vgl. Abschnitt 9.4.7) eingebunden. Es werden alle möglichen Kombinationen von `A` und `N` mit mindestens 2 und höchstens 5 Formen abgedeckt, aufgrund des Fokus auf nominale MWEs mit Ausnahme rein adjektivischer Verbindungen.

Abbildung 16 zeigt exemplarisch mehrere Pfade mit Kongruenzprüfungen sowie einen generischen Pfad für Formen ohne Morphologieangaben.

2 Main Items: Zweiteilige MWEs kommen am häufigsten vor und dienen oft als „Bausteine“ für längere Ausdrücke, sie gliedern in in Nomenkomposita `NN` (inkl. (NPN) sowie MWE mit post- bzw. pränominalem Adjektiv, vgl. Tabelle 10. Längere Kandidaten werden in LEMoN nicht rekursiv aufgebaut, sondern als explizite Muster bis 5 MIs modelliert - dies erhöht zwar den Wartungsaufwand, ermöglicht aber auch eine feinere Kontrolle. Die Deaktivierung des Graphen `AN` ohne alle Konstruktionen, die mit `AN` beginnen, ebenfalls zu deaktivieren ist ein Beispiel. Die Koordination läuft über `btwNN`. Einfügungen und Substitutionen bilden die übrigen Fälle ab.

Muster	Beispiel
NA	*agujero negro*
AN	*medio plazo*
NN	*año luz, campo de golf*

Tabelle 10: Nominale MWE mit 2 MI

Wie oben beschrieben ist der Graph AN in der Basiskonfiguration deaktiviert. Dieser Effekt wird in Ablationslauf 2 evaluiert, siehe Abschnitt 10.4.2.

3 Main Items: Dreiteilige Muster entstehen meist durch Modifikation oder durch Überkomposition. Koordinierte Sequenzen wie *drogas duras e ilegales* werden als eine MWE mit 3 MIs (NAA) erfasst, jedoch nur in dem Sonderfall, wenn sie über *e* koordiniert werden (vgl. Abschnitt 9.4.7).

Muster	Beispiel
NNA	*arma de destrucción masiva*
NAN	*escuela pública de arte*
ANA	*alto tribunal militar*
AAN	*primera gran guerra*
NNN	*centro de control de calidad*
NAA	*acción financiera arriesgada*

Tabelle 11: Nominale MWEs mit 3 MI

4 Main Items: Vierteilige Muster beruhen oft auf einem dreiteiligen Kern. Sie entstehen meist durch zusätzliche Modifikation wie ein weiteres Adjektiv oder durch Überkomposition, etwa indem ein weiterer N-de-N-Komplex angefügt wird.

Muster	Beispiel
NAAA	*tribunal militar internacional permanente*
ANNA	*alta corte de justicia penal*
NNNA	*sistema de gestión de calidad total*
ANNN	*Gran Premio de Motociclismo de España*

Tabelle 12: Nominale MWEs mit 4 MI

5 Main Items: Fünfteilige Muster sind selten. Wie [B. Daille 1994] festhält, sind die meisten nominalen MWEs zweigliedrig. Diese Beobachtung ist auf das Spanische

übertragbar. Größere Einheiten entstehen überwiegend durch Koordination sowie durch Modifikation und Überkomposition.

Muster	Beispiel
NANNA	*redes integrales de servicios de salud mental*
ANANA	*alta comisión independiente de derechos humanos*

Tabelle 13: Nominale MWEs mit 5 MI

9.4.7 Konnektoren

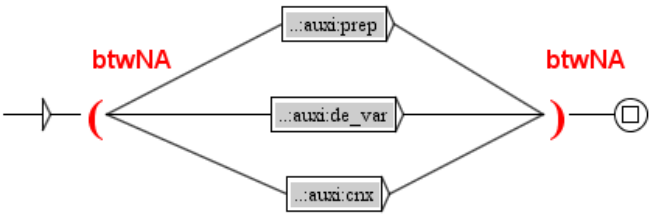

Abbildung 17: `btwNA.grf`

Zwischen den MIs können je nach Kombination zusätzliche Elemente stehen, etwa Determinierer oder Präpositionen. Diese werden über Konnektor-Graphen modelliert, die in den Kandidatengraphen zwischen den MIs eingebunden sind.

Es gibt die folgenden Konnektoren:

btwAA Elemente zwischen zwei Adjektiven

btwAN Elemente zwischen Adjektiv und Nomen

btwNA Elemente zwischen Nomen und Adjektiv

btwNN Elemente zwischen zwei Nomen

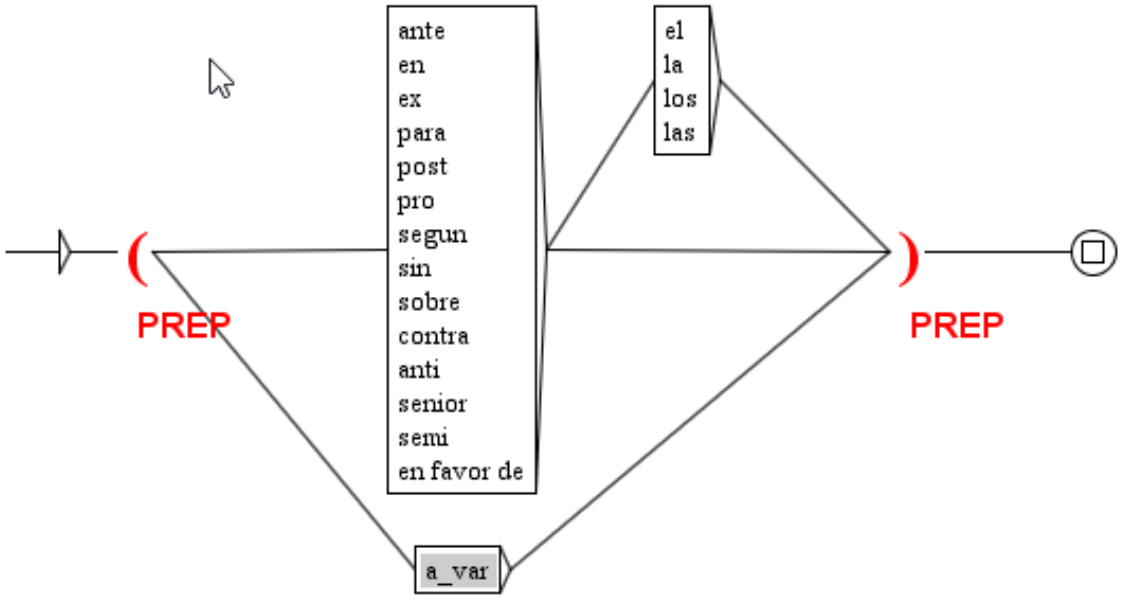

Abbildung 18: `prep.grf`

Der Konnektor `btwNA` in Abbildung 17 umfasst Elemente zwischen `N` und `A`. Dazu zählen Präpositionen (vgl. Abbildung 18), Varianten von `de` mit einfachem oder komplexem Determinierer (vgl. Abbildung 19), außerdem Leerzeichen, ein Bindestrich oder die Serienkonjunktion `e` (vgl. Abbildung 20).

Koordination (`e` vs. `y`): Standardmäßig ist nur `e` aktiv. `e` ist die Allomorphie von `y` vor vokalischem Anlaut. Mit `y` würden nahezu alle regulären Aufzählungen (`N y N`, `A y A`) durchkommen, was die Präzision stark senkt. Die Beschränkung auf `e` reduziert dieses Rauschen deutlich. Bei Bedarf kann `y` zugeschaltet werden. Die Evaluation in dieser Arbeit basiert auf der Konfiguration mit `e` ohne `y`.

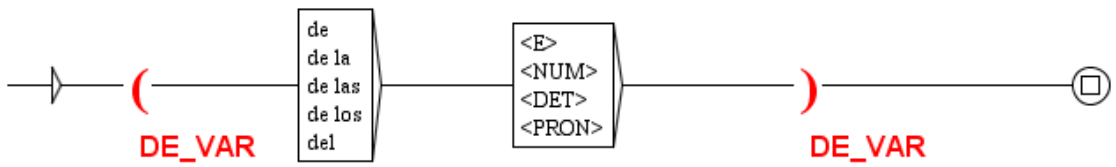

Abbildung 19: `de_var.grf`

Bindestriche sind im Spanischen seltener als im Französischen, aber starke Indi-

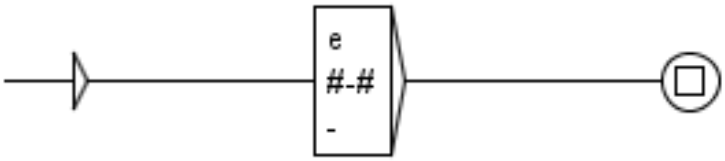

Abbildung 20: `cnx.grf`

katoren für MWEs. In allen syntaktischen Graphen bindet `cnx.grf` den Bindestrich standardmäßig als Konnektor zwischen allen MIs ein. Die Notation `# -#` erzwingt, dass kein Leerzeichen vor oder nach dem Bindestrich steht. Das folgt der spanischen Orthografie, übersieht aber gelegentliche Tippfehler im Korpus. Eine großzügigere Behandlung würde die Fehlertreffer stark erhöhen, weil der Gedankenstrich oft mit demselben Zeichen gesetzt ist. Schreibvarianten (Bindestrich/Leerzeichen) werden downstream im DELA-Export über das Gleichheitszeichen `=` abgebildet (z. B. `Castilla=La=Mancha`).

Zusammenfassung: Die syntaktische Steuerung begrenzt die Erkennung auf Wortfolgen, die sich zwischen einem Begrenzungszeichenpaar oder nach einem Determinierer befinden, und koppelt diese mit explizit modellierten Kandidatenmustern für 2 bis 5 MI. Konnektor-Hilfsgraphen erfassen reguläre Einfügungen (z. B. *de*, Bindestrich, *e*), während atypische Strukturen bewusst deaktiviert bleiben und nur in der Ablation ausgewertet werden. So maximiert LEMoN den Recall bei kontrollierbarer Präzision und liefert über den Transducer-Output valides DELA-Format.

10 LEMoN-Evaluation

Whatever you are not changing you are choosing.

Laurie Buchanan

Dieses Kapitel führt die Fäden zusammen. Auf Basis der in Kapitel 9 entwickelten lokalen Grammatiken wende ich die LEMoN-Pipeline (Kapitel 8) auf die in Kapitel 7 beschriebenen Korpora und Wörterbücher an. Die Pipeline erzeugt zunächst Konkordanzen, deren Gesamtausgabe ich als *Index* bezeichne. Aus dem Index leite ich durch Deduplikation ohne Berücksichtigung von POS das *Kandidateninventar* ab. Durch Deduplikation mit eindeutigen Lemma-POS-Muster-Paaren entsteht schließlich das Endprodukt `LEMoN.dic`.

Im Folgenden evaluiere ich die LEMoN-Pipeline und `LEMoN.dic` anhand eines Hauptlaufs und komplementärer Ablationsläufe.

Begriffsbestimmung:

- *Matches (Index)* alle im Hauptlauf gefundenen MWEs, einschließlich Doppelungen (Beleginstanzen).
- *Kandidatentypen (Kandidateninventar)* deduplizierte Matches ohne POS-Berücksichtigung (oberflächenidentische MWEs zählen einmal).
- *Lemma-POS-Paare (`LEMoN.dic`)* deduplizierte Einträge mit eindeutiger Kombination aus Lemma und POS-Muster.

10.1 Ablauf und Fragestellungen

Die Evaluation der LEMoN-Pipeline erfolgt in zwei komplementären Schritten. Erstens wird im Hauptlauf das Gesamtergebnis betrachtet: Die Analyse richtet sich hier auf den Inhalt und die Qualität von `LEMoN.dic`, also auf Fragen wie Präzision, Längenprofile und Musterverteilung. Zweitens folgt eine systematische Ablationsstudie auf einem Teilkorpus. Sie zielt nicht primär auf das Inventar, sondern auf die Pipeline selbst. Durch gezieltes An- und Abschalten einzelner Parameter werden die Effekte von Designentscheidungen in den LEMoN-Grammatiken quantifiziert. Damit lassen sich Robustheit, Verzerrungen und die internen Mechanismen der Pipeline sichtbar machen.

Die konkreten Fragestellungen für diese beiden Untersuchungen sind:

I) Fragestellungen zum Hauptlauf und LEMoN.dic:

I) Kandidateninventar und Abdeckung

1. Wie verteilt sich das Kandidateninventar über POS-Muster und MI-Längen?
2. Welche der modellierten POS-Muster sind im Korpus belegt?
3. Wie verteilen sich die Lesarten, also die Zahl der POS-Muster pro Lemma?
4. Welchen Anteil am Inventar tragen eindeutige gegenüber mehrdeutigen Lemmata?

II) Qualität und Präzision

1. Wie hoch ist die Präzision in einer Stichprobe von 50 zufällig gewählten Fällen für MWEs und reguläre NPs?
2. Lassen sich für alle erkannten POS-Muster akzeptable Belege im Wörterbuch finden?
3. Welche Fehlerklassen treten auf und wie lassen sie sich adressieren?

III) Interne Schachtelungen

1. Wie groß ist der Anteil verschachtelter MWEs und wie verteilt er sich über die Tokenlängen?
2. Wie unterscheidet sich die Tokenlängenverteilung eingebetteter MWEs gegenüber einbettenden MWEs?

IV) Lemmatisierung und Ambiguität

1. Wie verteilen sich die Ambiguitätsgrade, also die Zahl der POS-Muster pro Lemma?
2. Welchen Anteil am Inventar tragen eindeutige gegenüber mehrdeutigen Lemmata?
3. Welche POS-Muster treten häufig gemeinsam am selben Lemma auf?

V) Produktivität der Erstwörter

1. Welche Erstwörter sind produktiv, und welcher Anteil der Lemmata entfällt auf MWEs mit produktivem Erstwort gegenüber solchen ohne?

→ Mögliche Antworten, die sich aus den erhaltenen Daten ergeben, werden im Abschnitt 10.2 diskutiert.

Fragestellungen zu den Ablationsläufen:

1) Wie verteilt sich das Laufkandidateninventar über POS-Muster und MI-Längen?

2) Wie verschieben die Parameter `AN`, `syn_atyp` und `<E>` die MI-Längenverteilung?

3) Wie stabil sind die häufigsten POS-Muster (`NN`, `NA`, `NNN` ...) über die Läufe, gemessen an Rangordnung und Rangkorrelation?

4) Wie groß ist die inhaltliche Überlappung jedes Laufs mit der Baseline (Jaccard-Koeffizient der Lemma-POS-Paare)?

5) Welchen Einfluss hat die Index-Policy (`Longest Matches` vs. `All Matches`) auf verschachtelte MWEs und auf die Relation von Matches zu Kandidatentypen?

⟶ Mögliche Antworten, die sich aus den erhaltenen Daten ergeben, werden in der Beschreibung der einzelnen Läufe in Abschnitt 10.3 diskutiert.

Leitfragen zum Vergleich der Ablationsläufe:

1) Erhöht `AN` die Trefferqualität oder vor allem das Rauschen, und wie groß ist der Nettogewinn nach Deduplikation auf Lemma-POS-Ebene?

2) Sind `<E>`-Transitionen im Hinblick auf Abdeckung längerer MI-Ketten unverzichtbar?

3) Lässt sich der Verzicht auf `syn_atyp` rechtfertigen, gegeben deren Einfluss auf Verteilung und Vergleichbarkeit der Muster?

4) Welche Unterschiede entstehen durch `All Matches` gegenüber `Longest Matches` in Bezug auf Verschachtelungen, Arbeitsaufwand und Informationsgewinn?

5) Wie robust ist die Baseline-Konfiguration gegenüber den Alternativen insgesamt zu bewerten?

⟶ Mögliche Antworten, die sich aus den Daten ergeben, werden in Abschnitt 10.5 diskutiert.

10.2 Hauptlauf: Analyse von LEMoN.dic

Die im Rahmen des Hauptlaufs erzeugte DELA-Wörterbuchdatei `LEMoN.dic` bildet das zentrale Ergebnis der Pipeline und das Herzstück dieser Arbeit. Der folgende Auszug in Tabelle 14 zeigt die Form der Einträge im DELA-Format[44]. In `LEMoN.dic` wird die erkannte Einheit als Vollform gespeichert, das Feld für kanonische Form bleibt leer (zwischen Komma `,` und Punkt `.`). Da es sich um nominale MWEs handelt, ist jede Einheit als syntaktische Form als `.N` kodiert. Im Anschluss wird das Feature `MI` angegeben, das markiert, über welchen MI.Graph die konkrete Einheit erkannt worden ist.

Viele Oberflächenformen beginnen mit sehr produktiven Erstwörtern wie *acta*, *comunidad* oder *político*. Um diese Dominanz zu durchbrechen, habe ich einen beliebig gewählten alphabetischen Abschnitt genommen und dort die jeweils ersten zehn Einträge mit einem neuen Anfangswort ausgewählt. Als Ausnahme ist Zeile 11 zusätzlich aufgenommen: Sie folgt unmittelbar auf Zeile 10 und illustriert den Fall mehrerer konkurrierender POS-Muster für dasselbe Lemma.

#	Dateizeile	Eintrag (DELA)
1	1882572	`jurista González,.N+MI=NN`
2	1882614	`juristas arrestados,.N+MI=NA`
3	1882655	`jurídicas de aplicación del Tratado,.N+MI=ANN`
4	1882657	`jurídico en la dominación mundial,.N+MI=ANA`
5	1882658	`justa aplicación del Derecho de la competencia,.N+MI=ANAN`
6	1882758	`justas batallas libradas sobre las constituciones,.N+MI=AAAN`
7	1882795	`justeza de nuestra elección común,.N+MI=NNA`
8	1882798	`justicia -incluida,.N+MI=NA`
9	1883415	`justiciabilidad de los actos,.N+MI=NN`
10	1883419	`justicias en asuntos penales,.N+MI=NNA`
11	1883420	`justicias en asuntos penales,.N+MI=NNN`

Tabelle 14: Auszug aus `LEMoN.dic`

Eine kurze Analyse dieses Ausschnitts zeigt einige Charakteristika von `LEMoN.dic`:

- Eigennamen sind in den Lexika häufig als `N` annotiert. Das führt zu Erkennungen wie #1. Im Recall-orientierten Ansatz ist das vertretbar, für bestimmte Fälle jedoch fragwürdig. Daneben gibt es die Fehlerklasse klar falscher Annotationen sowie Wörter, die gar nicht im Lexikon erfasst sind. Die in Abschnitt 7.3

[44] `Oberflächenform,.N+MI=POS-MUSTER`, zur ausführlichen Dokumentation der DELA-Notation siehe Kapitel 6.

berichtete OOV-Rate von etwa 19 je 100 k Token ist daher relevant: Lücken in den Basislexika begünstigen Fehlmatches und abgebrochene verschachtelte Ketten.

- Wie beschrieben ist jeweils nur die erste Zeile eines neuen Anfangsworts aufgenommen. Wir haben das semantische Feld *Jura* in diesen zehn Sprüngen nicht verlassen, doch zwischen #1 und #11 liegen über 848 Zeilen. Das erklärt sich durch zwei Mechanismen: die hohe Produktivität bestimmter Felder und zahlreiche mehrfach erkannte Lemmata mit abweichenden Lesarten.

- Die Auswahl der MI-Muster ist für einen kurzen Ausschnitt variantenreich: vier Kandidaten mit 2 MI, vier mit 3 MI und zwei mit 4 MI. 5 MI treten hier nicht auf. Das entspricht der Erwartung, da trotz der *Longest Matches*-Policy kurze Ketten dominieren und längere seltener sind.

- Mehrfachzeilen zur selben Oberfläche wie in #10 und #11 entstehen, wenn verschiedene POS-Muster auf dieselbe Wortkette passen: `justicias en asuntos penales` als `NNA` und `NNN`. Die Lesart `NNA` ist korrekt, denn *penales* bezieht sich adjektivisch auf *asuntos* und meint hier „Strafsachen“. Als Nomen bezeichnet *penales* hingegen „Strafvollzugsanstalten“. Solche Varianten sind für breite, recall-orientierte Grammatiken in Kauf zu nehmen. Eine spätere Kuratierung kann redundante oder wenig plausible Lesarten reduzieren.

- Das Beispiel #6 ist im Lexikon insgesamt viermal vertreten. In der hier gezeigten ersten Instanz ist es als `AAAN` ausgezeichnet, dabei handelt es sich um eine Fehlkennung: *batallas* wurde fälschlich als Adjektiv in einem Lexikon hinterlegt.

- Viele Stellen lassen sich gezielt nachschärfen, etwa die Bindestrichform in #8. Das ist typisch für Bootstrapping: Mit jeder Iteration legen Korpus und Lexikon neue Verbesserungsmöglichkeiten frei. Der hier dokumentierte Stand zeigt, wie dieser Prozess gesteuert werden kann. Weitere Optimierungen sind möglich und hängen vom Zielkorpus und den eingesetzten Lexika ab. Ein Wechsel der Ressourcen verändert die Datenbasis und erfordert neue Anpassungen.

Darüber hinaus dient dieser Ausschnitt nur der Illustration. Die inhaltliche Auswertung erfolgt unten auf Basis der Stichprobe.

10.2.1 Kandidateninventar und Abdeckung

Leitfragen

- Wie verteilt sich das Kandidateninventar über POS-Muster und MI-Längen?
- Welche der modellierten POS-Muster sind im Korpus belegt?
- Wie verteilen sich die Lesarten, also die Zahl der POS-Muster pro Lemma?
- Welchen Anteil am Inventar tragen eindeutige gegenüber mehrdeutigen Lemmata?

Die Kernzahlen von `LEMoN.dic` sind in Tabelle 15 aufgeführt. Die Gesamtzahl von 7 095 273 Matches beinhaltet mehrfache identische Treffer und schrumpft durch Deduplikation auf 1 544 240 gefundene Formen. Das ergibt eine Dichte von 2 533 Matches je 100 000 EuroParl-Token. Gemessen an der Größe des EuroParl-Korpus (vgl. Tabelle 6) mag diese Zahl zunächst gering wirken. Das erklärt sich durch die Index-Einstellung *Longest Matches*, die kürzere MI-Konstellationen zugunsten längerer Spannen überdeckt. Der Effekt ist in Ablationslauf L6 (siehe Abschnitt 10.4.7) quantifiziert. Auch hier sei auf den Effekt der OOV-Wörter verwiesen, die den Recall direkt betreffen.

Tabelle 15: Kernzahlen von `LEMoN.dic`

Metrik	**Anzahl**	**Anteil**
Matches (*Index*, Beleginstanzen gesamt)	7 095 273	
Kandidatentypen (*Kandidateninventar*)	1 544 240	100 %
Lemma-POS-Paare (`LEMoN.dic`)	3 772 683	—
Lemmata mit genau 1 POS	583 041	37,76 %
Lemmata mit >1 POS	961 199	62,24 %
Ø POS pro Lemma	2,44	—
Unterschiedliche POS-Muster (Typen)	55	—

Von diesen 1 544 240 Lemma-Typen haben rund 38 % nur eine einzige Lesart in Form eines POS-Musters, die verbliebenen rund 62 % haben zwei oder mehr POS-Muster. Für LEMoN relevant ist die Größe der eindeutigen Lemma-POS-Paare, denn diese

stellen das Inventar dar: `LEMoN.dic` beinhaltet 3 772 683 Lemma-POS-Paare. Die durchschnittliche Zahl von POS-Mustern pro Lemma beträgt 2,4. Von den im Graphen modellierten 55 möglichen POS-Mustern sind im Hauptlauf alle 55 beobachtet worden. Die theoretische Obergrenze ohne Nebenbedingungen läge bei 60 Mustern für Längen 2 bis 5 über dem Alphabet `N,A`. Durch die Bedingung „mindestens ein N“ entfallen die reinen A-Ketten, und durch die bewusste Deaktivierung von `AN` verbleiben 55 modellierte Muster. Der Effekt der Entscheidung, `AN` zu deaktivieren, wird in der Ablation (Lauf L2, Abschnitt 10.4.2) gezeigt.

Abbildung 21 zeigt die Verteilung der 55 POS-Muster. Das Bild ist stark rechtsschief. Wenige Klassen treten sehr häufig auf, viele sind selten. Die häufigsten Muster decken den Großteil der Token ab. Danach fällt die Kurve über 2 und 3 MI-Kandidaten in eine lange, dünn besetzte Randzone, endend in erwartbar seltenen Ketten mit 4 oder 5 MI.

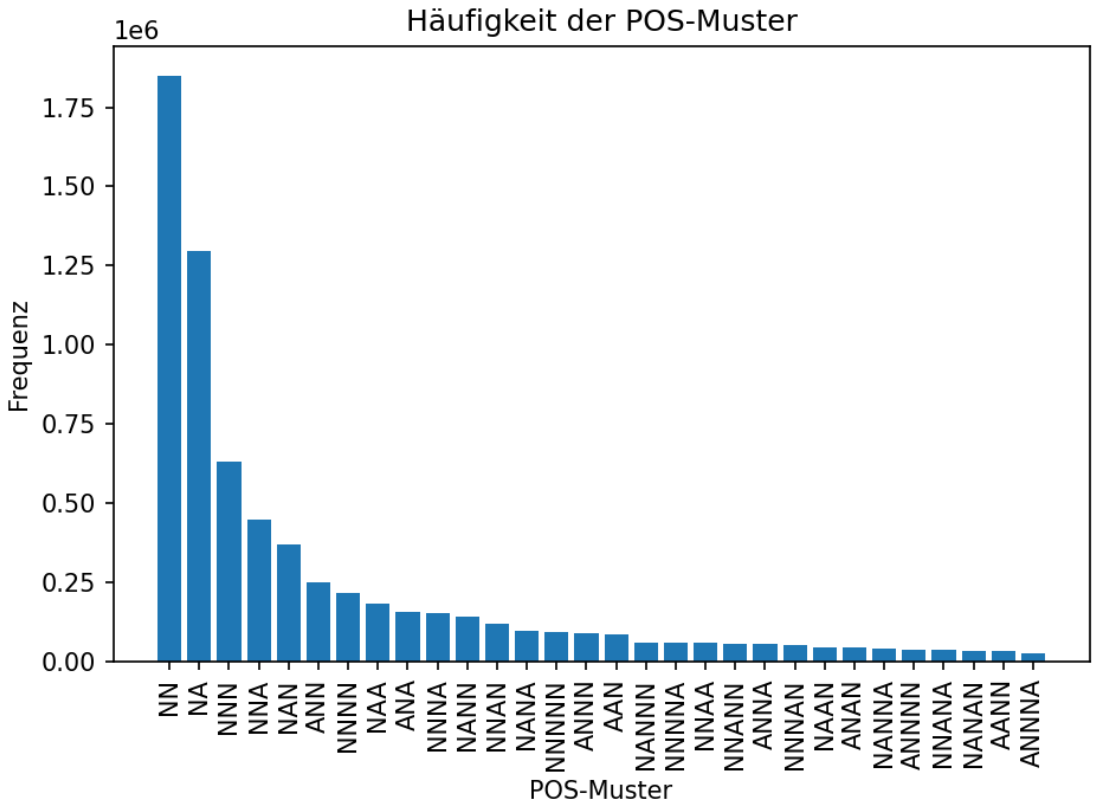

Abbildung 21: Häufigkeit der POS-Muster

Die 10 häufigsten POS-Muster sind in Tabelle 16 aufgeführt, eine vollständige Auflistung der einzelnen POS-Muster und ihrer Häufigkeiten in EuroParl findet

sich im Anhang B, dazu Beispielbelege. Erwartungsgemäß überwiegen die nominal geprägten Muster: Alle vier rein nominalen Muster (`NN`, `NNN`, `NNNN`, `NNNNN`) sind in den Top 14 vertreten. Die ersten drei stehen in den Top 7. Das erste Muster mit mehr als einem `A` folgt auf Platz 8 (`NAA`). Das erste Muster mit drei `A` steht auf Platz 35 (`NAAA`), die vier Muster mit vier `A` belegen die letzten vier Plätze (`ANAAA`, `AAANA`, `NAAAA`, `AAAAN`). Ein gestapeltes Längenprofil ist in Abbildung 22 zu sehen. Es veranschaulicht, wie sich die einzelnen POS-Muster je MI-Länge auf die Gesamtzahl der gemessenen Muster verteilen.

POS-Muster	**Häufigkeit**	**Anteil (%)**	**kumuliert (%)**
`NN`	1 850 406	26,08	26,08
`NA`	1 297 115	18,29	44,37
`NNN`	629 250	8,87	53,24
`NNA`	448 201	6,32	59,56
`NAN`	370 479	5,22	64,78
`ANN`	248 112	3,50	68,28
`NNNN`	216 892	3,06	71,34
`NAA`	182 730	2,58	73,92
`ANA`	155 765	2,20	76,12
`NNNA`	150 654	2,12	78,22

Tabelle 16: Top-10 POS-Muster nach Häufigkeit

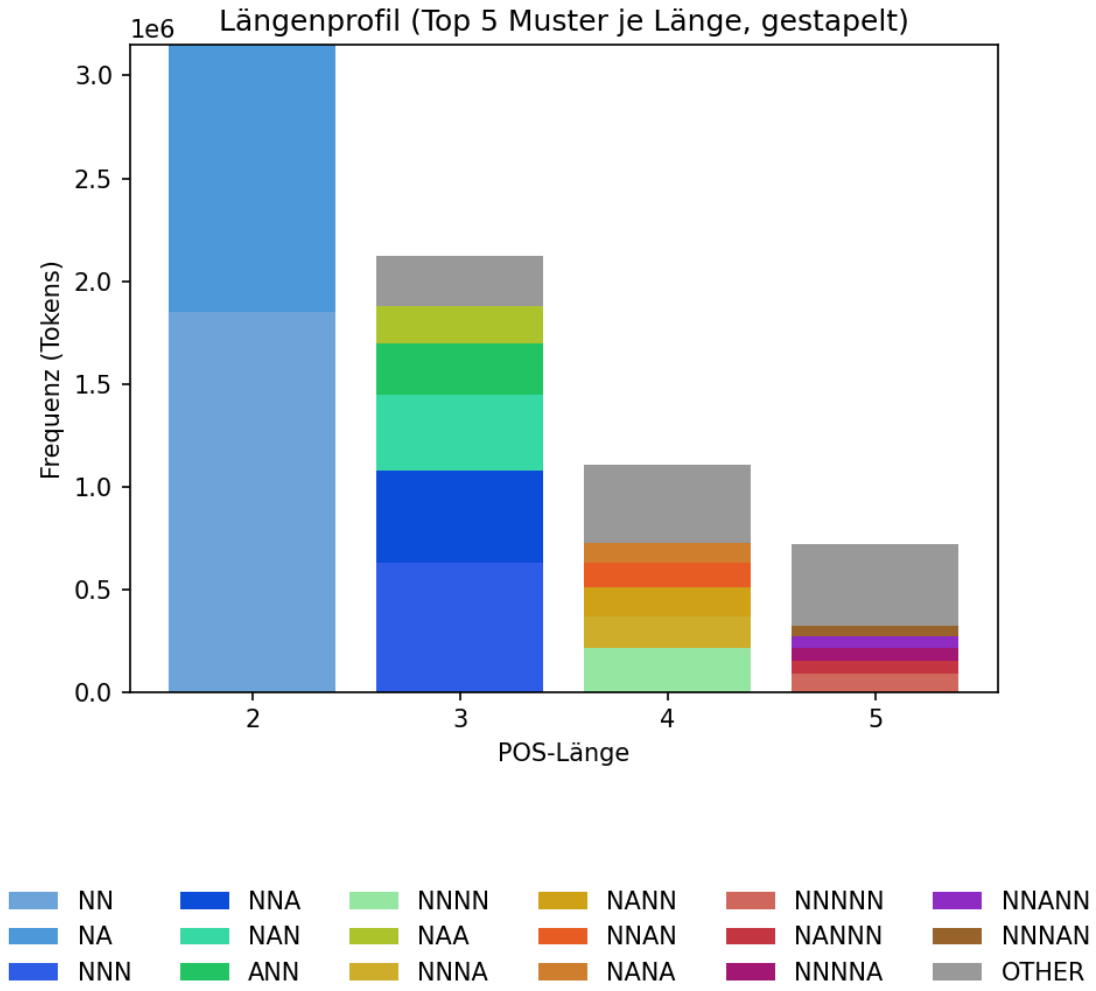

Abbildung 22: Gestapeltes Längenprofil der POS-Muster

10.2.2 Qualität und Präzision

Leitfragen

- Wie hoch ist die Präzision in einer Stichprobe von 50 zufällig gewählten Fällen für MWEs und reguläre NPs?
- Lassen sich für alle erkannten POS-Muster akzeptable Belege im Wörterbuch finden?
- Welche Fehlerklassen treten auf und wie lassen sie sich adressieren?

Für die Beantwortung dieser Leitfragen habe ich aus `LEMoN.dic` eine Zufallsstichprobe von N=50 Kandidaten ohne Zurücklegen gezogen. Die Ziehung erfolgte mit einem Python-Skript unter Verwendung eines festen Seeds (`-seed 20241028`) und der integrierten Funktion `random.sample()`. Die Datei wurde zuvor nach Oberflächenformen dedupliziert. Die Bewertung der Stichprobe ist manuell erfolgt und dient der Kalibrierung der Pipeline. Sie ist illustrativ und erhebt keinen Anspruch auf Vollständigkeit.

Tabelle 17: Zufallsstichprobe aus `LEMoN.dic`

ID	Kandidat	Kategorie
1	**interés constructivo**	MWE
2	**organización juvenil**	MWE
3	caso de un canadiense	NP
4	posible **ejemplo negativo**	NP+
5	buena **gestión financiera** de nuestros fondos	NP+
6	ayudantes de los **medios de comunicación**	NP+
7	**nacionalistas extremistas** del **ala derecha**	MWE+
8	posiciones viriles	NP
9	**desarrollos científicos** sobre los efectos	NP+
10	**consecuencias humanas** de esta problemática	NP+
11	situación de Kiruna	NP
12	**derechos intrínsecos del niño**	MWE+
13	**dignidad humana** en el centro de la construcción de la Unión	NP+
14	desperdicio irresponsable de nuestros recursos	NP
15	desarrollo del **mercado único** digital	NP+
16	campo del **capital riesgo**	NP+
17	Globalización -solicitud	NO
18	aplicación urgente de esta **ayuda comunitaria**	NP+
19	**proceso de valoración** de las conclusiones relativas	NP+
20	referéndum en Dinamarca sobre la participación en el euro	NP
21	misión de **observación electoral** de la **Unión Europea** en Liberia	NP+
22	delimitación clara de los **residuos domésticos**	NP+
23	**producto para la salud**	MWE
24	mujeres en la casa	NP
25	campo de las medicinas necesarias	NP
26	aplicación de las regulaciones de la **hora de verano**	NP+
27	creación de **instalaciones de almacenamiento**	NP
28	millones de agricultores en la Comunidad	NP
29	buen **clima laboral**	NP+
30	**dietas de asistencia parlamentaria**	MWE+
31	análisis de la situación jurídica	NP
32	tema político básico	NP
33	peticiones de estas personas	NP
34	concentración en el **tema estrella** de los últimos meses	NP+
35	**estrella de David**	MWE
36	aplicación del plan de recuperación del **atún rojo**	NP+
37	organizaciones de las **minorías étnicas**	MWE

ID	Kandidat	Kategorie
38	laboratorios de investigación del **mundo occidental**	NP+
39	fantasma de la **inmigración cero**	NP+
40	**zona judicial europea común**	MWE
41	médicos extranjeros de **Médicos Sin Fronteras**	NP+
42	**Comité Ejecutivo Nacional**	MWE
43	función de la mujer	NP
44	**cláusula de saneamiento**	MWE
45	**régimen draconiano**	MWE
46	tercera intervención	NP
47	**crecimiento económico** día	NO+
48	plétora de estructuras existentes	NP
49	uno de los **países miembros de la UE**	NP+
50	celebración de dichos acuerdos	NP

Analyse der Stichprobe: Für die Auswertung unterscheide ich die folgenden Klassen, der Definition von MWEs aus Abschnitt 2.1.3 folgend:

- MWE Die Zielklasse der nominalen MWE: eine polylexikale, syntaktisch zusammenhängende Nominaleinheit mit mindestens lexikalisiertem (gegebenenfalls idiomatischem) Verhalten. = **9** Treffer
- MWE+ Die gesamte Oberfläche stellt eine MWE dar, die intern mindestens eine weitere MWE enthält. = **3** Treffer
- NP Regulär gebildete NP ohne MWE-Status der Gesamtphrase (im Rahmen von LEMoN akzeptiert). = **16** Treffer
- NP+ Regulär gebildete NP, die mindestens eine nominale MWE als eingebettetes Teilsegment enthält (die Gesamtphrase selbst ist dabei keine MWE). = **20** Treffer
- NO Keine der obigen Kategorien (z. B. Artefakt, unvollständige oder fehlerhafte Spanne, Koordinationsrest). = **1** Treffer
- NO+ Es handelt sich um keine reguläre NP, enthält aber mindestens eine MWE. = **1** Treffer

Daraus ergibt sich folgendes Bild: 24 % der Stichprobentreffer sind korrekt erkannte MWEs (MWE oder MWE+, 12 von 50). Rechnet man die für LEMoN akzeptierten regulär gebildeten NP hinzu, ergibt sich ein Anteil von 56 % (28 von 50). Allerdings enthalten die meisten Treffer mindestens eine MWE (66 %, 33 von 50).

16 Kandidaten sind bloße NP ohne eingebettete MWE. Zusammengerechnet stellen MWEs, MWE-haltige und NP den überwiegenden Teil der Treffer dar, lediglich 1 Treffer der 50er Stichprobe enthält weder MWEs noch eine reguläre NP-Sequenz (NO).

Die Präzision berechne ich im Sinne des Ziels der LEMoN-Grammatiken entsprechend sowohl über die Zahl der MWEs insgesamt als auch über die Zahl der Matches, die mindestens eine MWE enthalten und zuletzt über die Zahl der fachlich vertretbaren nominalen Treffer. NO+ zählt in dieser Rechnung zwar zur Präsenz einer MWE, jedoch nicht zu den akzeptablen nominalen Treffern, NO entspricht Fehlmatches.

Präzision:

1) Präzision(MWE) = Anteil MWE (MWE, MWE+) an allen 50 Fällen
2) Präzision(Vorkommen) = Anteil der Klassen mit mindestens einer MWE (MWE, MWE+, NP+, NO+)
3) Präzision(Akzeptabel) = Anteil fachlich vertretbarer nominaler Treffer (MWE, NP, NP+)

Kennzahl	**Zähler/Nenner**	**Anteil**	**95%-CI (Wilson)**
Präzision(MWE)	12/50	24 %	[14,3, 37,4] %
Präzision(Vorkommen)	33/50	66 %	[52,2, 77,6] %
Präzision(Akzeptabel)	48/50	96 %	[86,5, 98,9] %
Fehlmatches (NO)	1/50	2 %	[0,4, 10,5] %

Tabelle 18: Präzisionssicht auf die Stichprobe

Interpretation:

- Präzision(MWE) = 24 %: etwa jede vierte zufällig gezogene Zeile ist eine einfache oder geschachtelte MWE.
- Präzision(Vorkommen) = 66 %: in rund zwei von drei Zeilen kommt mindestens eine MWE vor.
- Präzision(akzeptabel) = 96 %: fast alle Treffer sind vertretbare nominale Einheiten.
- Fehlmatches(NO) = 2 %: relativ selten, aber absolut $\approx 77\,000$ fehlerhafte Einträge.

Belegsuche zu POS-Mustern und Befund: Ergänzend zur Stichprobe habe ich bewusst qualitativ und illustrativ Belege für alle POS-Muster in `LEMoN.dic` gesucht[45] und in Anhang B dokumentiert. Da `LEMoN.dic` nicht normalisiert ist, folgen Groß- und Kleinschreibung der Korpusoberfläche. Die Belege sind manuell aus dem Wörterbuch ausgewählt und nach Eignung kuratiert. Wo möglich habe ich eindeutige einfache oder geschachtelte MWEs vorgezogen. Bei seltenen POS-Mustern habe ich auch zweifelhafte Fälle und reguläre NPs akzeptiert.

Zentraler Befund ist, dass für 52 von 55 modellierten Mustern jeweils mindestens ein sinnvoller Beleg vorliegt. Das ist angesichts modellierter Muster, die auch unwahrscheinliche Kombinationen beinhalten, bemerkenswert. Im Spanischen sind pränominale Adjektive selten, insbesondere gehäuft. Insofern sind die POS-Muster mit vielen führenden Adjektiven nicht mit der Erwartung modelliert worden, besonders ertragreich zu sein. Erwartungsgemäß finden sich lediglich für die Fälle mit mehr als zwei führenden Adjektiven (`AAAN`, `AAANA` und `AAAAN`) keine überzeugenden Belege. Es gibt jedoch Belege für `NAAAN` (*Agenda internacional específica para la próxima década*), `AAANN` (*próxima e inminente nueva propuesta de modificación*) sowie eine grenzwertige, sloganhafte MWE für `NAAAA` (*Autoridad europea alimentaria eficaz e independiente*). Das stützt die Einschätzung, dass sehr lange pränominale Adjektivketten vorkommen können, auch wenn sie extrem selten sind.

Aus der Durchsicht ergibt sich zudem ein klarer Eindruck zur inhaltlichen Qualität von `LEMoN.dic`. Die überwiegende Zahl der Treffer enthält MWEs oder NPs. Dominant sind jedoch viele Lesarten pro Treffer, die als *false positives* zählen. Hauptursachen sind tatsächliche Ambiguitäten zwischen Adjektiven und Nomen sowie Fehlauszeichnungen im Lexikon. Das deckt sich mit dem insgesamt rechtsschiefen Musterprofil und der ausgeprägten Lesartenvielfalt (Ø 2,4 POS-Muster pro Lemma).

Fehlerklassen: In der Stichprobe finden sich einige systematische Fehler, die ich in der nachfolgenden Übersicht zu Fehlerklassen zusammengefasst habe.

Präpositionalphrasen mit en Die Präposition *en* erscheint in den Stichproben ausschließlich an der rechten Grenze nominaler Gruppen und markiert in der Regel eine externe Verortung oder Rahmung. In den Fällen *en el centro* (ID 13), *en Dinamarca [...] en el euro* (ID 20), *en Liberia* (ID 21) sowie *en el tema* schließt die en-Präpositionalphrase (PP) eine vorhergehende NP oder MWE ab, ohne selbst

[45] mittels regulärer Ausdrücke auf den POS-Mustern in `LEMoN.dic` und anschließender manueller Sichtung der entsprechenden DELA-Zeilen.

Teil der Einheit zu sein. Der einzige Grenzfall ist *mujeres en la casa* (ID 24). Auch hier liegt funktional eine Verortung vor, eventuell kommt aber eine Lesart als MWE in Betracht, ähnlich der festen Wendung *ama de casa*.

Mögliche Maßnahme: Im Graph `prep.grf` ist *en* aktuell als möglicher Übergang zwischen allen vier Kombinationen (AN, NA, NN, AA) definiert. Zu prüfen wäre, ob es Übergänge gibt, bei denen *en* grundsätzlich nicht in die MWE-Spanne einbezogen werden sollte, um die Abgrenzung nominaler Kerne zu stabilisieren.

Quantifizierende Köpfe Einige nominale Treffer beginnen mit Zähl- oder Mengenangaben, die syntaktisch als Determinierer fungieren. Beispiele sind *millones de* (ID 27), *uno de los* (ID 48) und eventuell *tercera* (ID 44). In diesen Fällen steht der eigentliche Nominalkopf im nachfolgenden Komplement, nicht in der linken Position.

Mögliche Maßnahme: Diese Strukturen sollten idealerweise über die Graphen für komplexe Determinierer in `DET_ALL.grf` abgefangen werden. Möglich ist entweder ein Lückenbereich in den bestehenden Graphen oder eine fehlerhafte Pfadaktivierung. Eine Erweiterung der Determinierer-Graphen könnte solche quantifizierenden Voranstellungen explizit modellieren und aus der nominalen Spanne ausnehmen.

Generische Substantive Mehrere Treffer zeigen eine Nomen+Präposition-Konstruktion, die eher eine semantische Rahmung als einen Bestandteil der MWE bildet. Dazu gehören *caso de* (ID 3), *campo del* (ID 16), *análisis de* (ID 30), *creación de* (ID 26) oder *concentración en* (ID 33). Diese Strukturen sind syntaktisch korrekt, dienen aber primär als Einleitung kontextueller Nominalphrasen und nicht als Teil einer lexikalisierten Einheit.

Mögliche Maßnahme: Einführung einer Klasse „Rahmennomen“ für häufige generische Köpfe (neben den genannten etwa auch *situación, proceso, tema*), die analog zu `DET_ALL.grf` im Vorfeld der MWE platziert werden und als Marker, nicht als Bestandteil einer MWE gelten. Dadurch ließen sich syntaktische Einleitungsphrasen von echten Nominalkernen abgrenzen.

Attributive Adjektive ohne Bindung Häufige Adjektive wie *bueno*, *claro* oder *básico* treten produktiv in regulären NP-Mustern auf, aber ohne idiomatische Bindung. Beispiele sind *buena gestión financiera de nuestros fondos* (ID 5), *buen clima laboral* (ID 28), *tema político básico* (ID 31) und *dichos acuerdos* (ID 49).

Mögliche Maßnahme: Da schwache Adjektive typischerweise an den Rändern nominaler Gruppen auftreten, könnten sie graphenseitig über Kontexte gefiltert werden. Eine kleine Negativliste häufig generischer Adjektive wäre ausreichend, um de-

ren automatische Einbeziehung in MWE-Spannen zu verhindern.

Eigennamen und Toponyme Die Entscheidung, Eigennamen als Nomen zu behandeln, erhöht die Abdeckung im Parsing, führt aber zu Überspannungen. In der Regel sind Eigennamen nicht Teil einer MWE, sondern befinden sich im weiteren nominalen Kontext. Beispiele sind *situación de Kiruna* (ID 11), *referéndum en Dinamarca sobre la participación en el euro* (ID 20) oder *misión de observación electoral de la Unión Europea en Liberia* (ID 21). Der eigentliche Nominalkern liegt jeweils vor dem Eigennamen. Durch dessen Einbeziehung entsteht keine neue lexikalische Einheit, sondern eine referenzielle Erweiterung.

Mögliche Maßnahme: Eigennamen sollten nicht mehr als `N` annotiert werden, sondern etwa als NOMBRE. Entsprechend dieser Beobachtung könnte NOMBRE dann beispielsweise als rechter Kontext und zusätzlicher Marker modelliert werden.

Bindestriche und Satztrennzeichen In einem Treffer, *Globalización -solicitud* (ID 17), tritt der Bindestrich als Trennzeichen zwischen syntaktisch unabhängigen Segmenten auf, wird im Parsing jedoch als Bestandteil der Tokenverbindung interpretiert. Solche Fälle entstehen typischerweise durch Zeilenumbrüche, Parenthesen oder Interpunktionsfehler in der Quelltranskription. Der Bindestrich fungiert dabei nicht als Kompositionssignal, sondern als syntaktische oder typographische Markierung zwischen selbstständigen Einheiten.

Mögliche Maßnahme: Unter Umständen lassen sich die `btw`-Graphen optimieren hinsichtlich der Frage, welche Form von Bindestrichen (echte Bindestriche, Minus und Gedanken- bzw. Bereichsstriche) sie als Konnektoren akzeptieren. Eine weitere Möglichkeit, dieses Problem zu addressieren, könnte es auch sein, die Satzenderkennung in dieser Hinsicht zu optimieren.

Die beschriebenen Fehlerklassen deuten überwiegend auf Phänomene hin, die sich durch eine Anpassung des linken oder rechten Kontexts in den LEMoN-Grammatiken adressieren lassen. Auf diese Weise lässt sich mutmaßlich die Zahl der MWEs, die sich innerhalb einer zu langen Spanne befindet, zugunsten reiner MWE-Treffer reduzieren.

10.2.3 Interne Schachtelungen

Leitfragen

- Wie groß ist der Anteil verschachtelter MWEs und wie verteilt er sich über die Tokenlängen?
- Wie unterscheidet sich die Tokenlängenverteilung eingebetteter MWEs gegenüber einbettenden MWEs?

Geschachtelte MWEs sind MWEs, die als ununterbrochene Teilstrings in längeren MWEs auftreten, der englische Terminus ist „Nesting“. Der Einsatz der *Longest-Match*-Policy verhindert zwar das Zersplittern von Treffern (wie bei *All-Matches*), verdeckt aber zugleich, welche längeren MWEs aus kürzeren Bausteinen bestehen. Im Folgenden untersuche ich, in welchem Umfang sich in `LEMoN.dic` solche Schachtelungen nachweisen lassen und welche strukturellen Tendenzen sich daraus ableiten. Dazu prüfe ich, welche MWEs („Short MWE“) in anderen MWEs („Long MWE“) eingebettet vorkommen. Die Entscheidung, nominale MWEs nach einer *Longest-Match*-Policy zu extrahieren, wird in Lauf L6 (siehe Abschnitt 10.4.7) gesondert überprüft.

Für diese Untersuchung betrachte ich zunächst die Längenverteilung des Inventars in `LEMoN.dic` sowie die Längen der MWEs, die als zusammenhängende Teilstrings in größeren MWEs auftreten. Danach untersuche ich, welche längeren MWEs als Wirtsformen für solche Verschachtelungen dienen. Abschließend identifiziere ich jene MWEs, die besonders produktiv als Teilstrings auftreten und damit die Bausteine des Inventars in `LEMoN.dic` bilden.

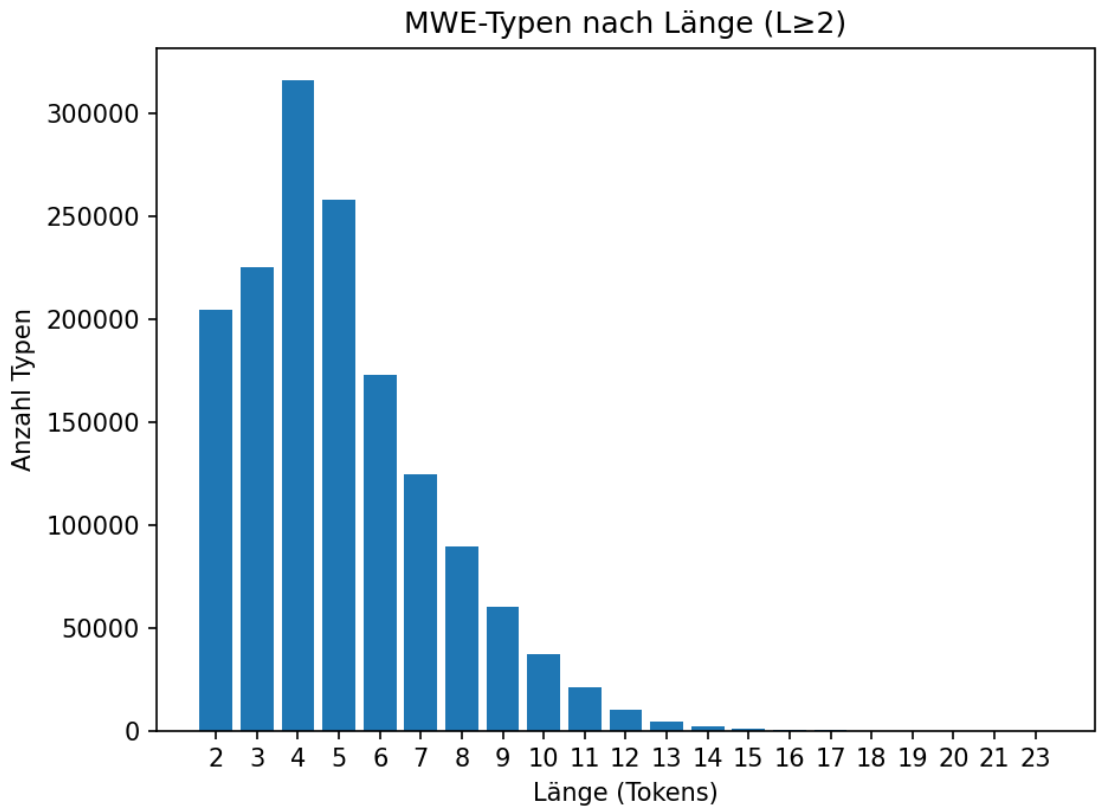

Abbildung 23: Typenzahl nominaler MWE nach Tokenlänge

Bisher bezog sich die Längenanalyse der MWEs vor allem auf die Zahl der MI. Für die Verschachtelungsanalyse werden die Einträge zusätzlich nach ihrer Tokenlänge betrachtet. Die meisten Einträge in `LEMoN.dic` haben eine Tokenlänge von vier (315 577), gefolgt von Typen der Längen fünf, drei und zwei. Ab sechs Token nimmt die Häufigkeit stetig ab. Im zweistelligen Bereich werden die Zahlen deutlich kleiner. Bei zwölf Token liegen noch rund 10 000 Typen vor. Danach dünnt der Ausläufer rasch aus, siehe Abbildung 23.

Wie beschrieben handelt es sich bei geschachtelten MWEs um solche, die als Teilstrings in größeren Einheiten vorkommen. Deren Längenprofil ist in Abbildung 24 dargestellt. Rund 40 Prozent der zweigliedrigen Typen kommen mindestens einmal als zusammenhängender Teilstring in längeren Einheiten vor. Drei-, Vier- und Fünf-Token-MWEs fallen gleichmäßig ab, beginnend bei etwa 29 Prozent über 21 Prozent bis zu etwa 11 Prozent. Danach flacht die Kurve zusehends ab. Ab einer Länge von acht Token liegt der Wert unter 3 Prozent. Spätestens hier ist in vielen Korpora eher mit Rauschen als mit tatsächlicher Verschachtelung zu rechnen. Im Fall von

EuroParl ist jedoch auch mit komplexen, mehrfach geschachtelten institutionellen Nominalphrasen wie *Comisión de Derechos de la Mujer e Igualdad de oportunidades* zu rechnen.

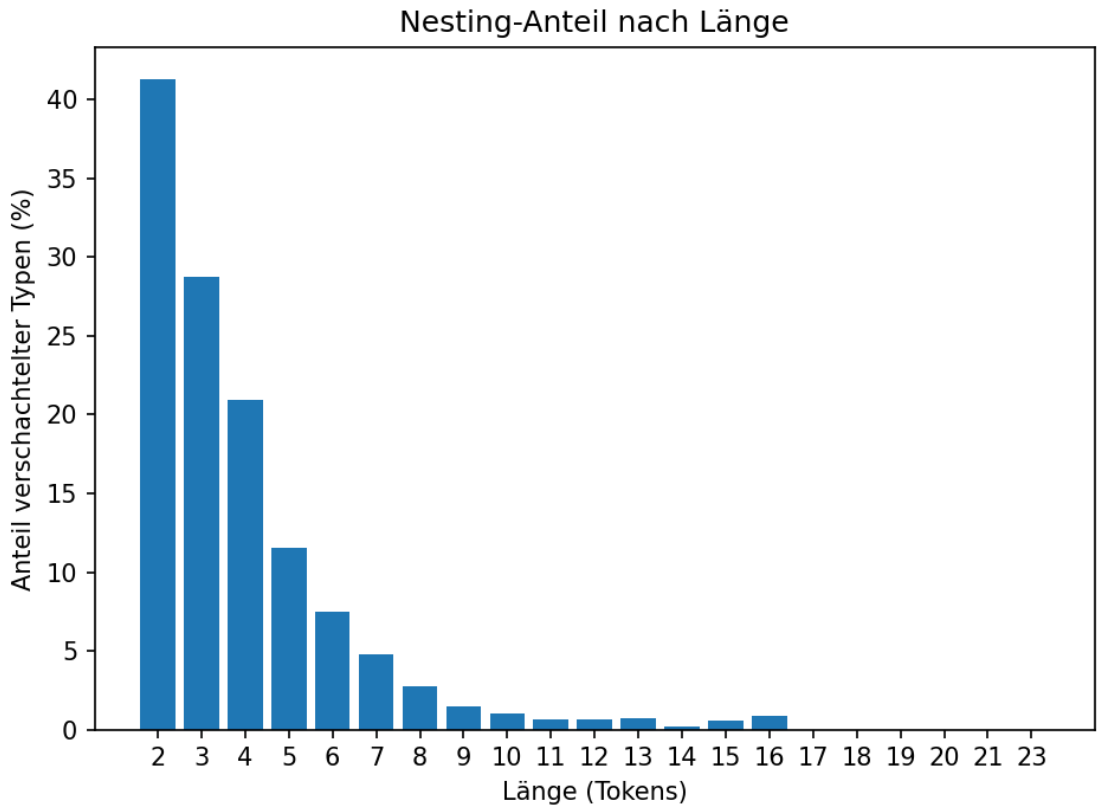

Abbildung 24: Anteil verschachtelter MWE je Tokenlänge

Das Längenprofil der MWEs, die verschachtelte MWEs beinhalten, ist in Abbildung 25 dargestellt und entspricht der Erwartung, dass mit steigender Länge die Wahrscheinlichkeit wächst, eine geschachtelte Einheit zu enthalten (blaue Balken). Drei- und Vier-Token-MWEs enthalten in rund 30 bzw. 43 Prozent der Fälle mindestens eine kürzere MWE. Ab fünf bis sechs Token steigt der Anteil auf über 80 Prozent und nähert sich rasch 100 Prozent an. Die orangefarbenen Balken zeigen den Anteil der Long-MWEs, die mindestens zwei kürzere MWEs als Teilstrings beinhalten. Dieser Anteil legt zwischen vier und sieben Token deutlich zu und steigt danach weiter an, bis ab etwa L=13 nahezu alle Long-MWEs mehr als eine kürzere MWE enthalten.

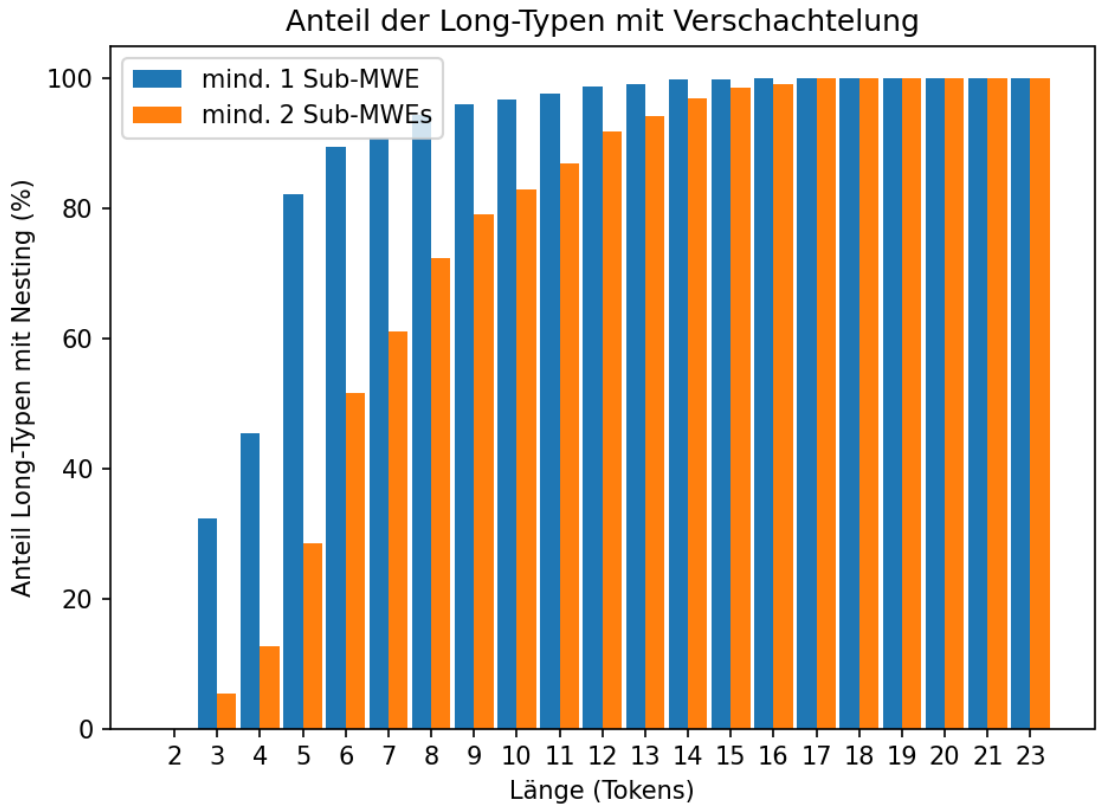

Abbildung 25: Längenprofil von Long-MWE, die Short-MWE einbetten

Die am häufigsten in Long-MWEs eingebetteten Short-MWEs sind in Abbildung 26 dargestellt. Die Liste wird der Domäne des Korpus entsprechend von politisch-institutionellen Begriffen dominiert. Diese Ausdrücke fungieren als Bausteine, die in zahlreichen MWEs und NPs eingesetzt werden. Ihre hohe Produktivität macht sie zu lexikalischen Knotenpunkten und unterstreicht die starke institutionelle Prägung des Korpus und insbesondere seines nominalen Registers.

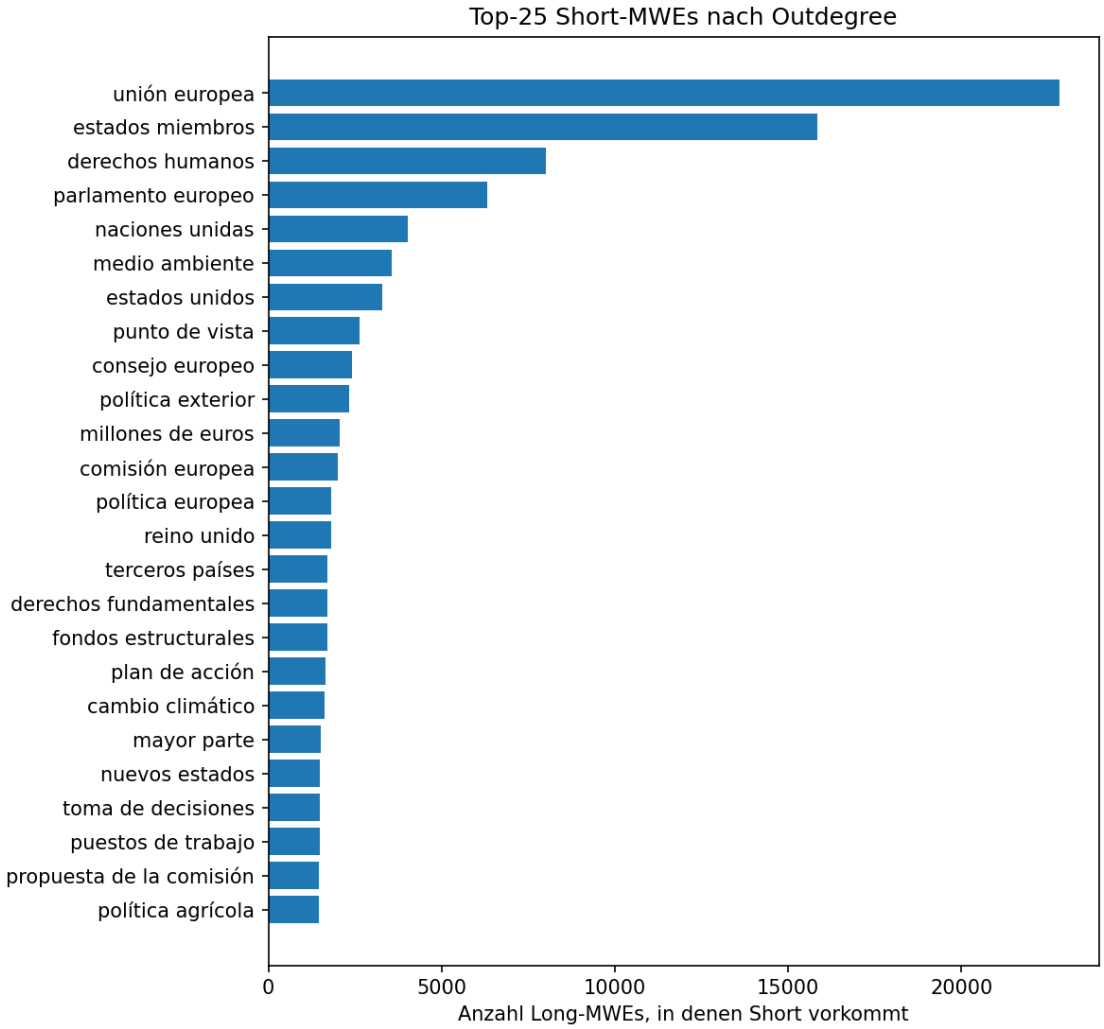

Abbildung 26: Die 25 am häufigsten eingebetteten Short-MWE (normalisiert)

10.2.4 Lemmatisierung und Ambiguität

Leitfragen

- Wie verteilen sich die Ambiguitätsgrade, also die Zahl der POS-Muster pro Lemma?
- Welchen Anteil am Inventar tragen eindeutige gegenüber mehrdeutigen Lemmata?
- Welche POS-Muster treten häufig gemeinsam am selben Lemma auf?

Viele Wörter sind in den Lexika mit mehr als einem POS ausgezeichnet. Dadurch entstehen mehrdeutige Interpretationen. Diese Ambiguität betrifft den Zielbereich von LEMoN. Im Folgenden untersuche ich, wie viele unterschiedliche POS-Muster sich auf einzelne Lemmata verteilen und welche Musterkombinationen häufig gemeinsam auftreten.

Abbildung 27 zeigt die Verteilung der Ambiguitätsgrade. Lemmata mit genau einem POS-Muster bilden die größte Gruppe. Gleichzeitig gibt es eine breite Zone mit zwei und drei Mustern sowie einen langen Ausläufer bis in hohe Werte. Ab etwa 15 bis 20 Mustern sinken die absoluten Zahlen deutlich, oberhalb von 30 Mustern treten nur noch Einzelfälle auf.

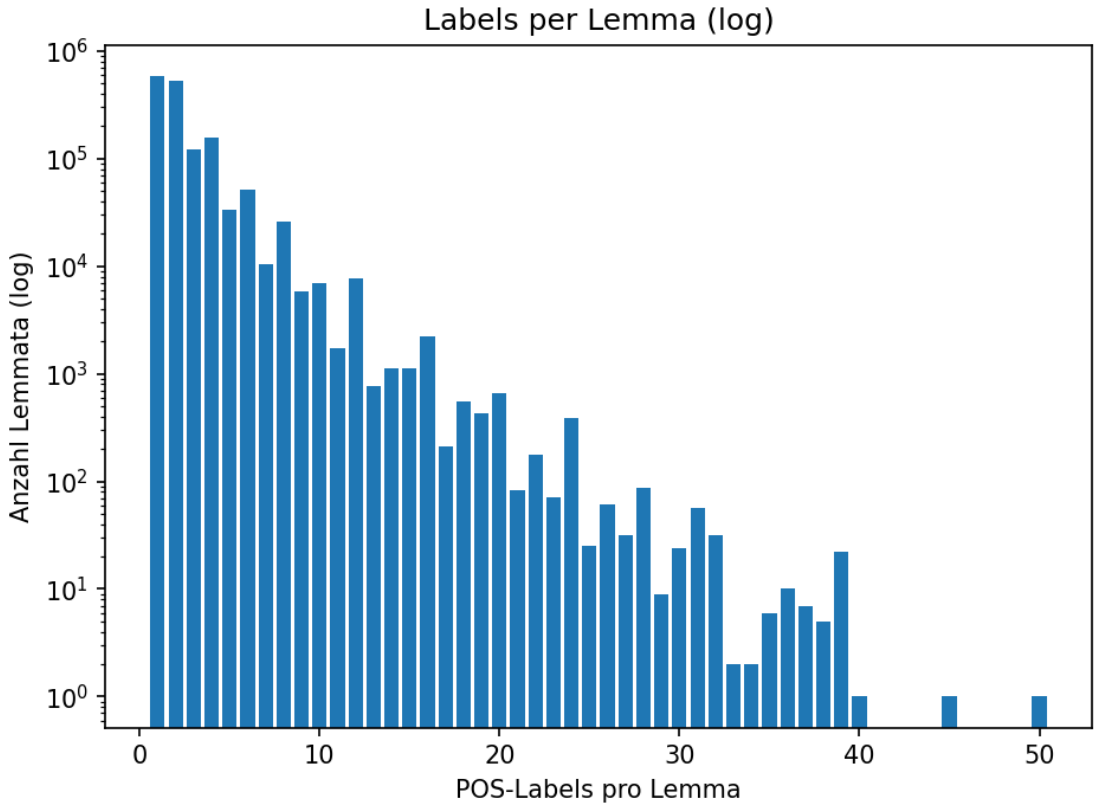

Abbildung 27: Lesarten pro Lemma

Das Boxplot in Abbildung 28 zeigt die Zahl der unterschiedlichen Lesarten pro Lemma, sortiert nach MI-Länge. Der Median ist durch die orangefarbene Linie markiert. Die Boxhöhe zeigt die überwiegende Menge der Lesarten innerhalb der jeweiligen MI-Länge. Die *Whisker* markieren seltene Ausreißer. Bei 2 MI liegt der Median bei eins. Ab 3 MI liegt der Median bei zwei. Lemmata mit 4 MI weisen die größ-

te Varianz auf. Das lässt sich dadurch erklären, dass bereits im Lexikon kodierte MWEs alternative kürzere Muster erlauben und damit zusätzliche Lesarten erzeugen. Lemmata mit 5 MI behalten den Median bei zwei, zugleich nimmt die Zahl hoher Ausreißer sichtbar ab. Dies deutet darauf hin, dass sich die möglichen Lesarten bei längeren Spannen stärker auf wenige Varianten bündeln. Gleichzeitig werden kürzere Alternativlesarten mit wachsender MI-Länge seltener, weil lange Spannen seltener vollständig modellierte Teilketten enthalten und die Randbedingungen der Graphen enger werden. Dadurch decken weniger kürzere Muster dieselbe Oberfläche noch ab.

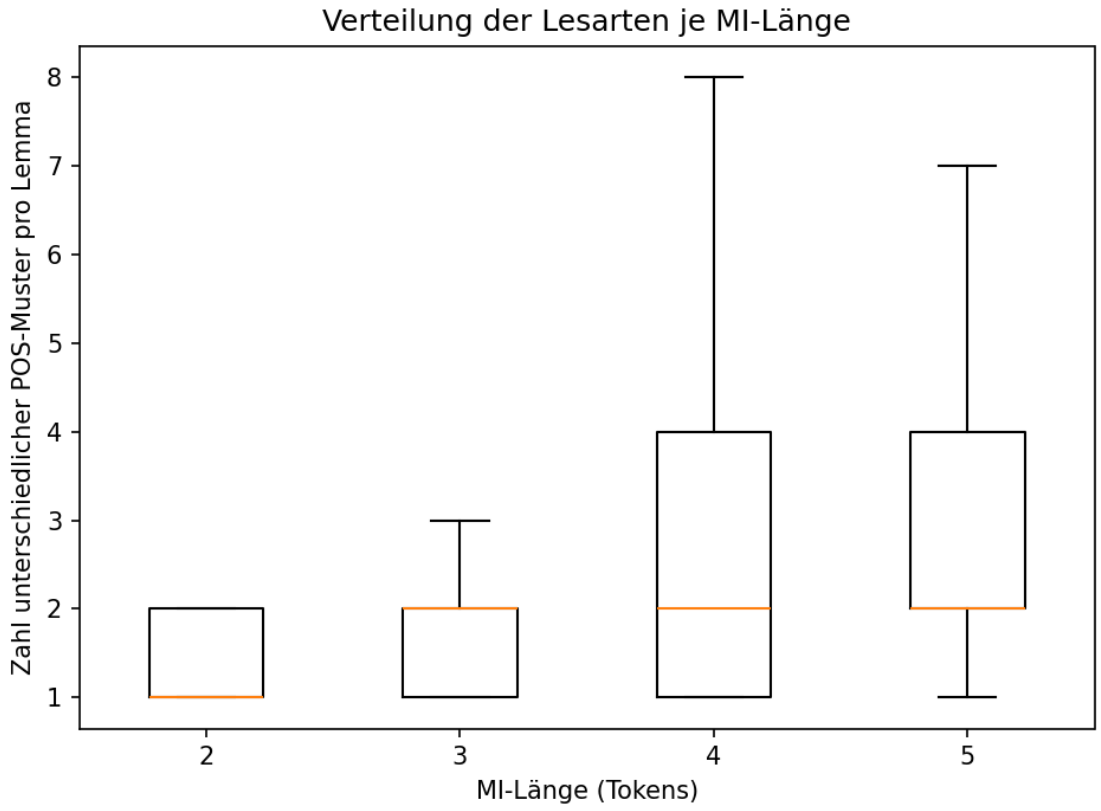

Abbildung 28: Verteilung der Lesarten pro Lemma je MI-Länge als Boxplots

In den lexikalischen Ressourcen ist bereits ein Bestand an MWEs kodiert, insbesondere im `ncomp_uab.dic` (vgl. Abschnitt 7.3). Dadurch kann ein Lemma sowohl in einer vollständig ausbuchstabierten Folge von MI erscheinen als auch in einer alternativen Lesart, in der Teilsegmente als `N` zusammengezogen sind. In der Folge kann dieselbe Oberfläche unterschiedliche MI-Längen aufweisen. Abbildung 29 zeigt, dass die meisten Lemmata nur in einer MI-Länge auftreten. Ein spürbarer Anteil besitzt

zusätzlich eine zweite Länge. Drei oder mehr Längen sind selten. Betroffen sind vor allem zweigliedrige und in geringerem Umfang dreigliedrige Einheiten.

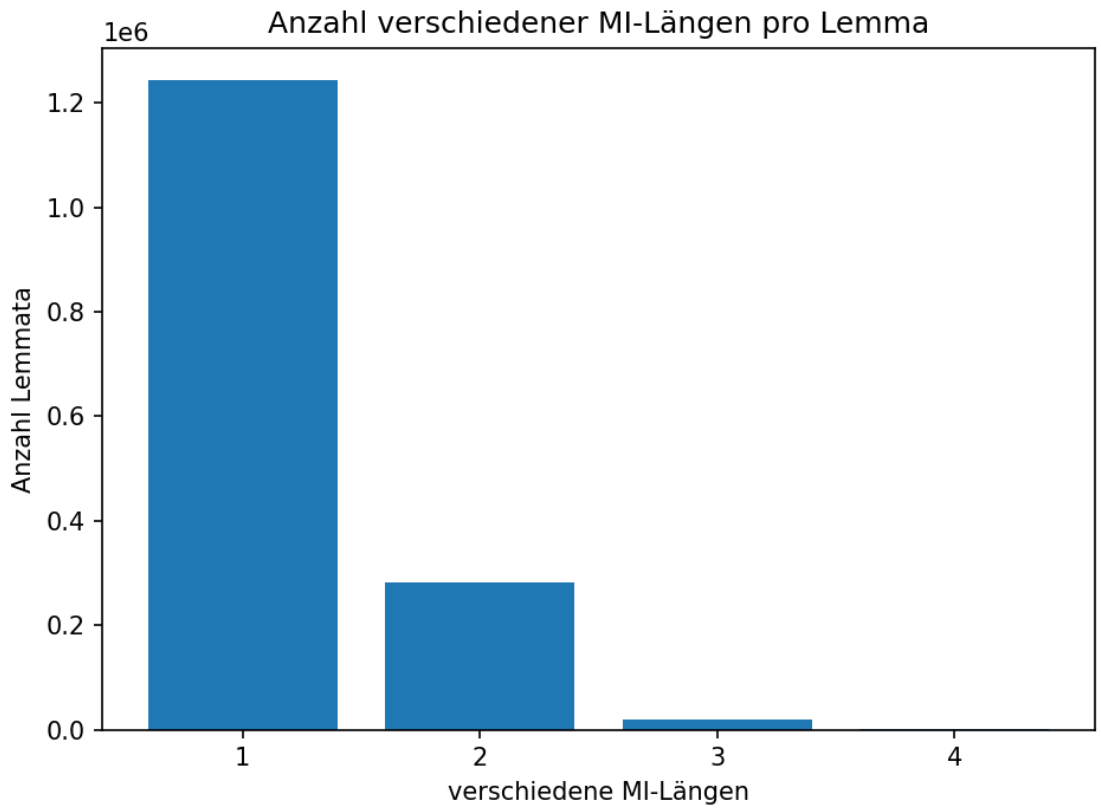

Abbildung 29: Anzahl unterschiedlicher MI-Längen pro Lemma

Abbildung 30 differenziert dieses Bild nach MI-Länge. Für jede Länge wird der Anteil der Lemmata gezeigt, die zusätzlich kürzere Varianten besitzen, längere Varianten besitzen oder insgesamt mehrere Längen aufweisen. Der Anteil der Lemmata mit mehreren Längen steigt besonders bei 4 MI an. Das ist konsistent damit, dass im Modell nur Längen von 2 bis 5 zugelassen sind. Lemmata mit 4 MI können sowohl kürzere als auch längere Alternativen haben. Lemmata mit 2 MI können nur längere Alternativen haben, Lemmata mit 5 MI nur kürzere.

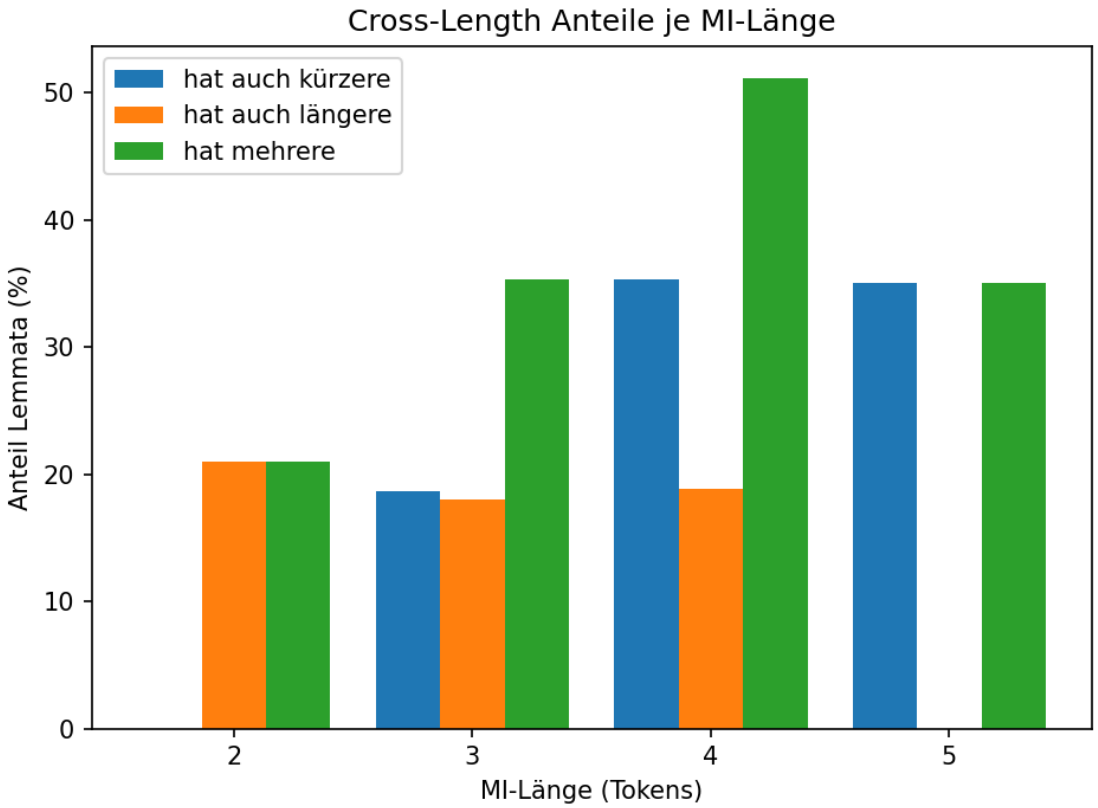

Abbildung 30: Anteile an Längenvarianz je MI-Länge

Von Interesse ist schließlich, welche POS-Muster konkret am selben Lemma miteinander konkurrieren. Die Heatmap in Abbildung 31 zeigt, dass es sich überwiegend um kleine Abweichungen handelt, zum Beispiel `NN` mit `NA` oder `NNN` mit `NNA`. Das spricht für nahe Alternativen am selben Lemma statt völlig unterschiedlicher Strukturen.

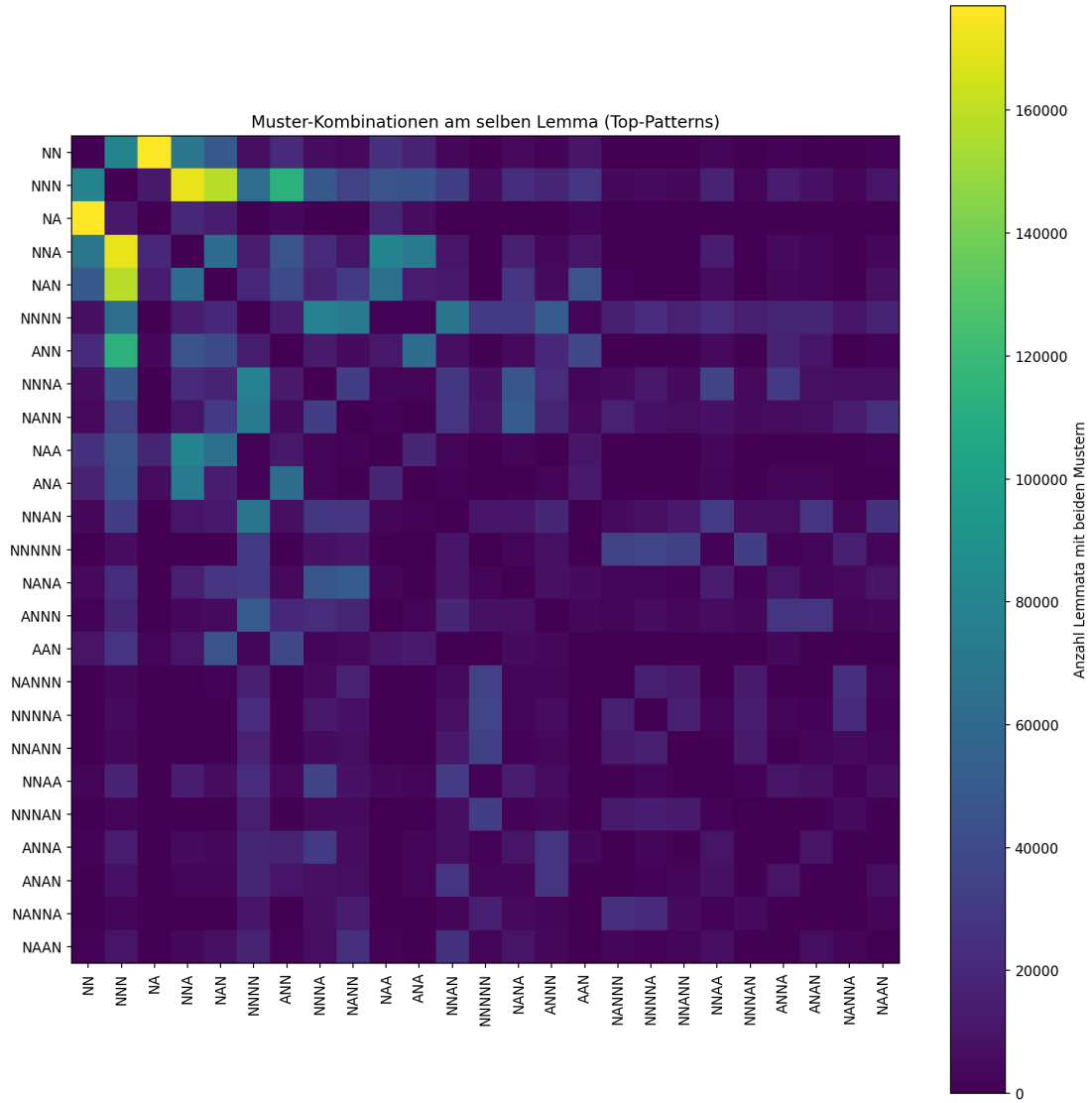

Abbildung 31: POS-Muster-Kombinationen am selben Lemma

Abschließend ein Beispiel, das mehrere der beschriebenen Phänomene bündelt. Der Ausdruck *red de banda ancha multimedia para la enseñanza* steht für *Breitband Multimedia Netz für die Lehre* und ist in sechs Lesarten identifiziert worden, kein ungewöhnlich hoher Wert:

```
red de banda ancha multimedia para la enseñanza,.N+MI=NAAAA
red de banda ancha multimedia para la enseñanza,.N+MI=NAAAN
red de banda ancha multimedia para la enseñanza,.N+MI=NNAA
red de banda ancha multimedia para la enseñanza,.N+MI=NNAAA
```

```
red de banda ancha multimedia para la enseñanza,.N+MI=NNAAN
red de banda ancha multimedia para la enseñanza,.N+MI=NNAN
```

Zwei Lesarten haben eine MI-Länge von 4, vier Lesarten haben eine MI-Länge von 5 MI. Zunächst: *red* ist korrekt als `N` erkannt worden, *multimedia* korrekt als `A`. Es liegt jedoch am rechten Rand eine Positionsambiguität vor, da *enseñanza* in den Ressourcen sowohl als `N` als auch als `A` erfasst ist. Letzteres ist vermutlich fehlerhaft, da der übliche Gebrauch nominal ist, wie in *material de enseñanza*. Zusammen führen die beiden Mechanismen zu Längenvarianten und zu `N` oder `A` an der letzten Position. Während es also durchaus auch plausible inhaltliche Gründe für mehrere Lesarten gibt, ist nicht zuletzt die Qualiät der verwendeten Lexika maßgeblich für die Menge von alternativen POS-Mustern.

Die kürzeren Lesarten entstehen plausibel dadurch, dass *banda ancha* bereits als MWE mit `N` markiert ist: Dies ist unter anderem im Wörterbuch `eswiki-alltitles-final.dic` der Fall. Dadurch kann *banda ancha* als ein `N` zusammengezogen werden, was die MI-Länge reduziert.

10.2.5 Produktivität der Erstwörter

Leitfrage

- Welche Erstwörter sind produktiv, und welcher Anteil der Lemmata entfällt auf MWEs mit produktivem Erstwort gegenüber solchen ohne?

Bereits ein schneller Blick in `LEMoN.dic` zeigt eine Charakteristik, die eng an das verwendete *EuroParl*-Korpus geknüpft ist. Es finden sich lange Blöcke aufeinanderfolgender Zeilen mit demselben Erstwort. Das spiegelt die hohe Dichte politisch fachlicher Terminologie, die dazu neigt, lange nominale Ketten zu bilden, sei es für Themen oder für Bezeichnungen von Personen, Institutionen und Organisationen. In Abschnitt 10.2.3 wurde gezeigt, dass dieses Inventar zu einem erheblichen Anteil aus Ketten kürzerer MWEs zusammengesetzt ist. Hier untersuche ich gezielt die Produktivität der ersten Wörter je Lemma.

Die Kurve in Abbildung 32 fällt gleichmäßig und ist stark rechtsschief. Das entspricht einem Verlauf, der einer Zipf-Verteilung ähnelt, mit wenigen sehr produktiven Erstwörtern. Die meisten Erstwörter sind hingegen selten oder sehr selten. Für die

Kuratierung von `LEMoN.dic` legt das zwei Pfade nahe. Erstens lohnt sich die Prüfung der hochproduktiven Musterträger und der Graphen, die sie generieren, weil gezielte Anpassungen hier einen großen Effekt haben können. Zweitens lohnt sich der Blick auf Wörter außerhalb dieser Produktivitätscluster, um zu klären, ob ihre Seltenheit auf Graphenführung oder Lexikonetikettierung beruht oder ob es sich um wenig produktive, aber korrekte Einheiten handelt.

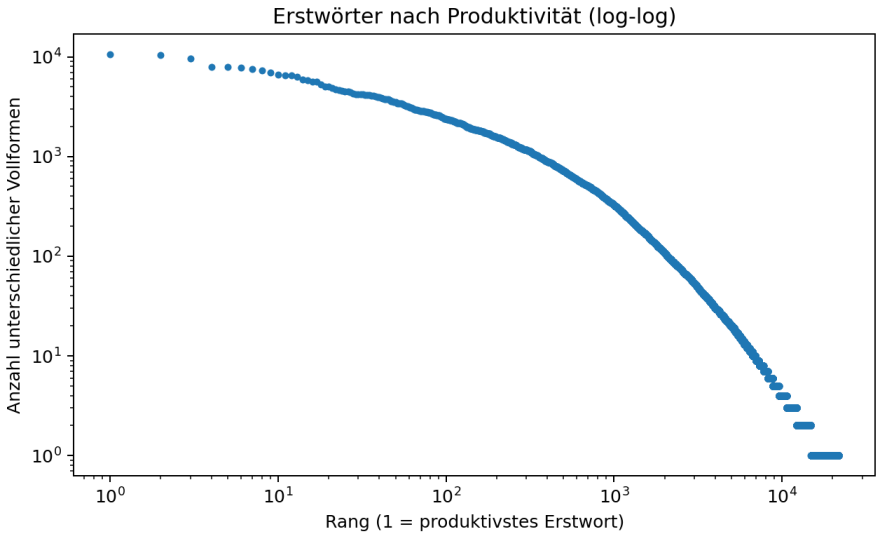

Abbildung 32: Erstwörter nach Produktivität

Die folgende Abbildung zeigt die häufigsten produktiven Erstwörter. An der Spitze stehen *política* und *mayor* mit jeweils über 10 000 unterschiedlichen Vollformen, dicht gefolgt von *cuestión*. Dominant sind generische, semantisch weite Nomen und Adjektive mit hoher Kombinatorik. Zur Illustration: *política* bildet über 10 000 Vollformen, etwa *política lenta*, *política Bush frente* oder *política vista de esta manera*. Ähnlich breit streuen *mayor* (*mayor Estado de derecho*, *mayor Unión Europea*) und *cuestión* (*cuestión Boeing*, *cuestión de fondo*).

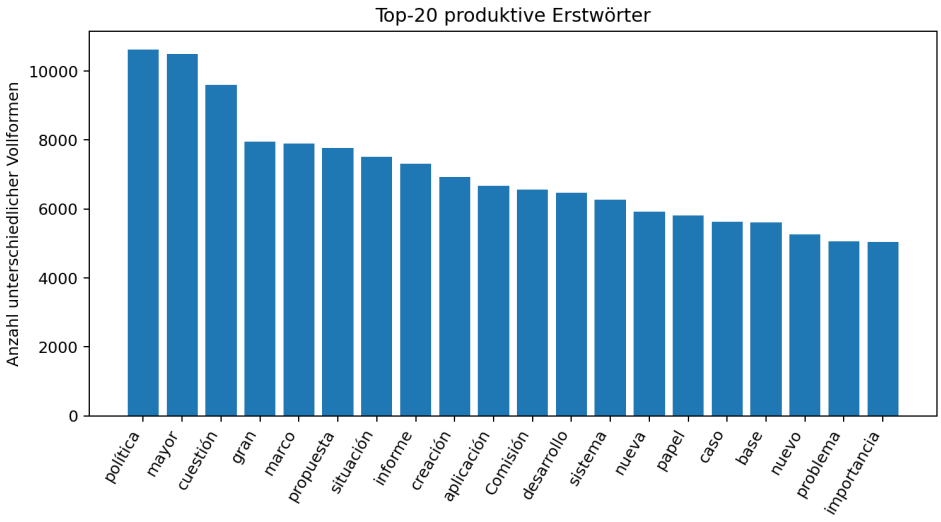

Abbildung 33: Top 20 produktive Erstwörter

Die Produktivität lässt sich auch als Abdeckung des deduplizierten Kandidateninventars ausdrücken. Abbildung 34 zeigt: Die zehn produktivsten Erstwörter decken etwa fünf Prozent der Lemmata ab. Die Top 20 etwa neun Prozent. Die Top 100 etwa ein Viertel. Die Top 1 000 rund drei Viertel. Das unterstreicht die starke Zentrierung des Inventars um wenige Musterträger.

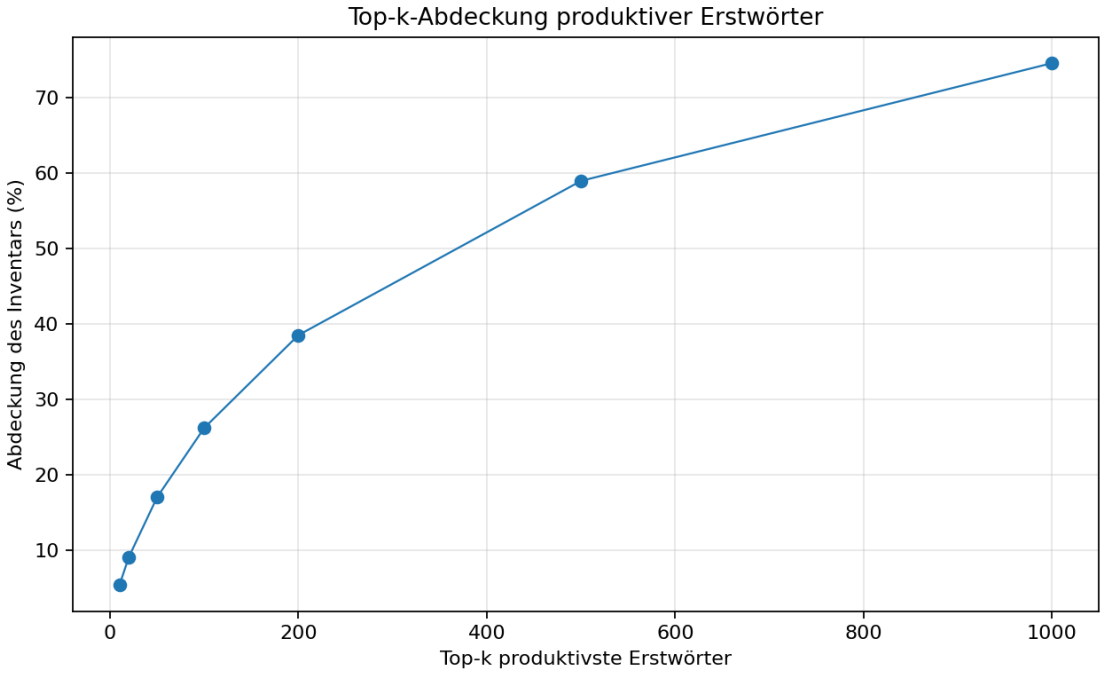

Abbildung 34: Lexikonabdeckung durch die produktivsten Erstwörter

10.2.6 Zusammenfassung

Damit beende ich die Analyse von `LEMoN.dic` und des Hauptlaufs. Die Auswertung hat ein breites Spektrum abgedeckt: Inventar und Abdeckung, Präzision und inhaltliche Qualität anhand einer Stichprobe mit Fehlerklassen, interne Schachtelungen von MWEs, alternative Lesarten und POS-Muster sowie die Produktivität der Erstwörter.

Aus den Befunden ergeben sich vier Kernaussagen. Erstens zeigt das Wörterbuch ein reiches, rechtsschief verteiltes Inventar, in dem wenige häufige Muster und viele seltene Muster koexistieren. Zweitens ist ein relevanter Anteil der Einträge intern verschachtelt. Kurze MWEs dienen oft als Bausteine längerer Einheiten, und mit wachsender Tokenlänge steigt die Wahrscheinlichkeit, mindestens eine eingebettete MWE zu enthalten. Drittens liegt auf Lemmaebene echte Ambiguität vor. Viele Lemmata tragen zwei oder mehr POS-Muster. Häufig konkurrieren nahe Musterpaare wie NN und NA oder NNN und NNA. Viertens prägen wenige sehr produktive Erstwörter große Teile des Inventars. Diese Köpfe sind semantisch weit und kombinatorisch flexibel und bilden sichtbar eine Ausprägung des spezifischen Korpus.

Praktische Schlussfolgerungen für die Kuratierung lassen sich klar priorisieren. Zuerst sollten hochproduktive Erstwörter und die sie tragenden Graphen geprüft werden, da kleine Anpassungen dort viele Einträge betreffen. Danach folgen Lemmata mit hohem Ambiguitätsgrad und wiederkehrenden Musterpaaren, um systematisch Mehrdeutigkeiten zu reduzieren. Ergänzend sind die in der Stichprobe identifizierten Fehlerklassen adressierbar, etwa Randpräpositionen, generische Rahmennomen, schwache adjektivische Modifikatoren und die Behandlung von Eigennamen.

Insgesamt zeigt sich ein eindeutiges Bild. Die Pipeline funktioniert und erzeugt ohne Kenntnis des zugrunde liegenden Korpus ein umfangreiches MWE-Inventar im DELA-Format, das bereits in vielen Fällen nutzbar ist. Gleichzeitig liegen genügend Ansatzpunkte für gezielte Verbesserungen vor, die von punktuellen Lexikon-Korrekturen über enger gefasste Kontexte in den Grammatiken bis zu kuratierten Listen produktiver Köpfe reichen. Bereits der Einsatz besser kuratierter, umfangreicher und auf das Zielkorpus zugeschnittener Wörterbücher wird die Qualität deutlich steigern.

Im nächsten Schritt wendet sich die Arbeit der Ablationsanalyse zu. Ziel ist es, die Pipeline selbst und ihre Designentscheidungen systematisch zu untersuchen und die im Hauptlauf beobachteten Effekte zu verifizieren und zu quantifizieren.

10.3 Ablationsstudie: Design und Durchführung

Leitfragen

1) Wie verteilt sich das Laufkandidateninventar über POS-Muster und MI-Längen?

2) Wie verschieben die Parameter `AN`, `syn_atyp` und `<E>` die MI-Längenverteilung?

3) Wie stabil sind die häufigsten POS-Muster (`NN`, `NA`, `NNN` ...) über die Läufe, gemessen an Rangordnung und Rangkorrelation?

4) Wie groß ist die inhaltliche Überlappung jedes Laufs mit der Baseline (Jaccard-Koeffizient der Lemma-POS-Paare)?

5) Welchen Einfluss hat die Index-Policy (`Longest Matches` vs. `All Matches`) auf verschachtelte MWEs und auf die Relation von Matches zu Kandidatentypen?

und

1) Erhöht `AN` die Trefferqualität oder vor allem das Rauschen, und wie groß ist der Nettogewinn nach Deduplikation auf Lemma-POS-Ebene?

2) Sind `<E>`-Transitionen im Hinblick auf Abdeckung längerer MI-Ketten unverzichtbar?

3) Lässt sich der Verzicht auf `syn_atyp` rechtfertigen, gegeben deren Einfluss auf Verteilung und Vergleichbarkeit der Muster?

4) Welche Unterschiede entstehen durch `All Matches` gegenüber `Longest Matches` in Bezug auf Verschachtelungen, Arbeitsaufwand und Informationsgewinn?

5) Wie robust ist die Baseline-Konfiguration gegenüber den Alternativen insgesamt zu bewerten?

In dieser Ablationsstudie messe ich die Wirkung einzelner Parameter, indem mehrere ansonsten identische Läufe ausgeführt werden. Die Parameter entsprechen möglichen konfigurativen Stellschrauben, die auch den Hauptlauf betreffen, dort sind sie in gleicher Weise gesetzt wie in Lauf L1. Die Datengrundlage ist jedoch nicht identisch. Im Hauptlauf kommt ein anderes Set von Default-Wörterbüchern zum Einsatz und in den Graphen wurden Korrekturen umgesetzt, die zum Zeitpunkt der Ablationsstudie noch nicht vorlagen. Die Effekte der Parameter sind innerhalb der Studie daher gut vergleichbar und lassen sich den jeweils veränderten Faktoren präzise zuordnen. Direkte Zahlenvergleiche mit dem Hauptlauf sind jedoch nicht zulässig, inhaltliche Bezüge lassen sich jedoch durchaus herstellen. Ziel ist es, die obigen Leitfragen zu beantworten und damit die Designentscheidungen beim Aufbau der LEMoN-Grammatiken zu bewerten.

In Tabelle 7.1 sind die Konfigurationen zusammengefasst, die in allen Läufen L1 bis L6 konstant gehalten wurden. Die spezifischen Abweichungen je Lauf sind anschließend dokumentiert.

Lauf	**AN**	**Atypische**	**<E>**	**Index**
L1 Baseline	–	–	+	Longest Matches
L2 AN aktiviert	+	–	+	Longest Matches
L3 Atypische aktiviert	–	+	+	Longest Matches
L3b AN + Atypische	+	+	+	Longest Matches
L4 <E> deaktiviert	–	–	–	Longest Matches
L5 AN + Atypische ohne <E>	+	+	–	Longest Matches
L6 Baseline + All Matches	–	–	+	All Matches

Tabelle 19: Übersicht der Ablationsläufe und ihrer Konfigurationen

10.3.1 Konfiguration

Die fixe Konfiguration entspricht derjenigen des Hauptlaufs (vgl. Abschnitt 7.1). Abweichend ist lediglich das verwendete Korpus: *EuroParl* v7 (ES), Teildatei `ep-07.txt` (77 413 kB). Die Wortlisten nach Vorverarbeitung umfassen 62 798 einfache lexikalische Einheiten, 12 255 komplexe Einheiten und 5 844 unbekannte Einheiten (alle Default-Lexika aktiviert, vgl. Abschnitt 7.3). Die spezifischen Faktoren, die in den Ablationsläufen variiert werden, sind im Folgenden jeweils beschrieben. Die bereinigten Transducer-Outputs wurden im DELA-Format exportiert und bilden lexikalische Kandidatenlisten.

10.3.2 Parameter

Die untersuchten Paramter spiegeln zentrale Designentscheidungen wider, die ich für den Hauptlauf getroffen habe. Die Ablation macht ihre Effekte transparent und nachvollziehbar:

AN Im Hauptgraph für 2 MI `2mi.grf` (vgl. Kapitel 9.4.6) habe ich den Graph für die Erkennung von `AN`-Kombinationen deaktiviert, da das extrem hohe Aufkommen regulärer Bildungen die manuelle Sichtung der Konkordanzen erheblich erschwert hat. Diese Entscheidung hat mutmaßlich den Recall reduziert und wird hier gezielt überprüft.

syn_atyp Basierend auf der Klassifizierung in [M. Mathieu-Colas 1996]) habe ich atypische syntaktische Kombinationen für spanische MWEs modelliert. Auch dieser Graph erzeugte starkes Rauschen, das die Analyse der Konkordanzen behinderte. Daher habe ich ihn im Hauptlauf deaktiviert, um ihn in der Ablation zumindest hinsichtlich seines numerischen Effekts zu prüfen.

`<E>` Die Graphen zwischen je 2 MI (`btwNA`, `btwNN`, `btwAN`, `btwAA`) erlauben leere Übergänge `<E>`, um auch Matches wie *año luz* zu erfassen. Diese „leeren" Durchgänge habe ich für diesen Lauf deaktiviert.

Index Überlappung der Matches für die Trefferaggregation in Unitex (mögliche Optionen: *Longest Matches*, *Shortest Match*, *All Matches*). Der Hauptlauf verwendet die Einstellung *Longest Matches* in der Annahme, dass damit auch geschachtelte MWEs erfasst werden. Diese Wahl unterdrückt jedoch vermutlich zahlreiche kürzere Varianten und interpretiert mitunter auch reguläre Bildungen oder Koordinationen als verschachtelte MWEs. Lauf L6 prüft diese Annahme mit der Alternative *All Matches*.

10.3.3 Varianten und Metriken

Zur Auswahl der gewählten Varianten:

1) Die Baseline L1 entspricht dem Hauptlauf, der sich aus dem Design der Grammatiken entwickelt hat.

2) Die Konfigurationen L2, L3, L3b, L4 und L5 prüfen die genannten Designentscheidungen.

3) L6 quantifiziert die Wahl der Index-Einstellung *Longest Matches* gegenüber *All Matches*.

4) Die partiellen Varianten mit deaktiviertem `<E>` wurden nicht realisiert, da die Wirkung bereits in L4 (reiner `<E>`-Effekt) und L5 (in Kombination mit beiden Faktoren) sichtbar wird.

Metriken: Die betrachteten Metriken sind eine Auswahl der im Hauptlauf verwendeten, sind aber ergänzt um den Jaccard-Koeffizienten:

Matches Alle gefundenen Matches nach der Index-Einstellung (s. o.), inklusive Doppelungen.

Kandidatentypen Alle gefundenen Matches, ohne Doppelungen (POS wird nicht berücksichtigt).

Lemma-POS-Paare Deduplizierte Einträge mit eindeutiger Kombination von Lemma und POS.

Jaccard Zur Messung der Überlappung zwischen den Läufen wurde der Jaccard-Koeffizient gegen die L1 Baseline berechnet (Schnittmenge der Lemma-POS-Kandidaten geteilt durch ihre Vereinigung).[46] Jaccard misst die inhaltliche Nähe der Lauf-Inventare (Lemma-POS-Ebene) zur Baseline. Hoch = ähnlich L1, niedrig = stark abweichend/ergänzend.

[46] berechnet mit A, B = Mengen der Lemma-POS-Kandidaten je Lauf als

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

10.4 Ablationsstudie: Läufe im Detail

10.4.1 Lauf 1: Baseline

Beschreibung. Dieser Lauf dient als ablationsinterne Baseline. Die Parameter entsprechen dem Hauptlauf aber Wörterbücher, Teile der Grammatiken und das Teilkorpus sind abweichend. L1 ist daher Referenz innerhalb der Ablation, dient aber nicht zur numerischen Gegenüberstellung mit dem Hauptlauf.

Konfiguration. `Atyp` = NEIN; `AN` = NEIN; `<E>` = JA; `Index` = LONGEST MATCHES.

Ergebnisse

Kernzahlen	
Matches (gesamt)	622 506
Dedupliziert	122 035
Lemma-POS	276 918
Jaccard vs. L1	1.000

Längenprofil	
MI=2	180 267
MI=3	226 051
MI=4	79 364
MI=5	136 824

Top-Muster	
NN	103 664
NA	76 603
NNN	71 472
NNA	52 574

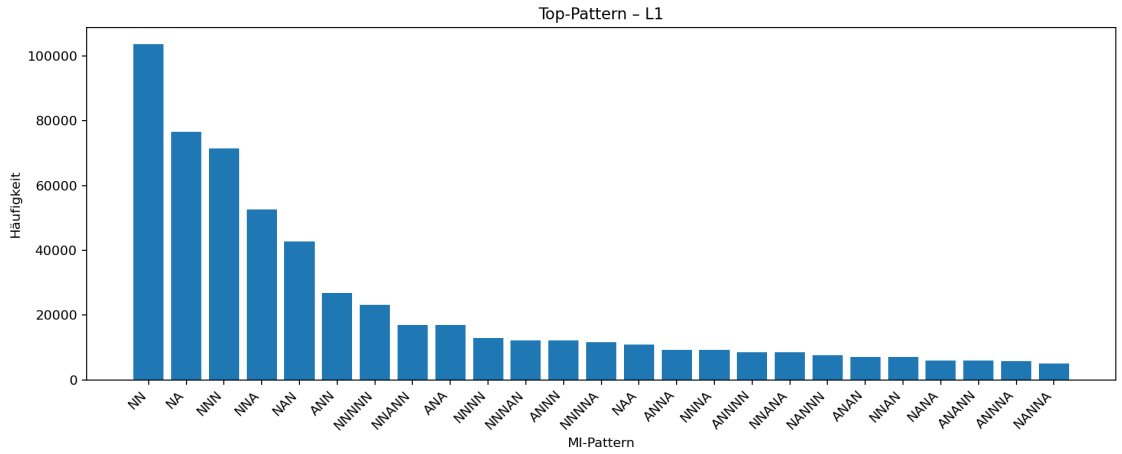

Abbildung 35: Top-Muster - L1: AN(–) Atyp(–) E(+) Longest Matches (Baseline)

Beobachtung. Die Baseline liefert ein stabiles Profil mit dominanten Mustern aus 2 bis 3 MI. Längere Sequenzen bilden den erwarteten Ausläufer. Das Fehlen von `AN` ist bewusst herbeigeführt. Die Top-Muster entsprechen bis Platz 6 dem Hauptlauf, ab Platz 7 zeigen sich Abweichungen. Das ist aufgrund der anderen Ressourcenlage bemerkenswert.

10.4.2 Lauf 2: AN-Graph aktiviert

Beschreibung. Im Hauptlauf wurde der `AN`-Graph (aus der Gruppe der 2-MI-Graphen) deaktiviert, da er subjektiv zu viel Rauschen erzeugte und die Ausdefinition lokaler Grammatiken behinderte. In diesem Lauf ist der `AN`-Graph aktiviert. Betroffen sind ausschließlich 2-MI-Kandidaten mit dem POS-Muster `AN`. Längere Sequenzen, die `AN` enthalten oder damit beginnen, werden unabhängig davon modelliert.

Konfiguration. `Atyp` = NEIN; `AN` = JA; `<E>` = JA; `Index` = LONGEST MATCHES.

Ergebnisse

Kernzahlen	
Matches (gesamt)	658 639
Dedupliziert	126 379
Lemma-POS	293 664
Jaccard vs. L1	0.943

Längenprofil	
MI=2	216 405
MI=3	226 051
MI=4	79 364
MI=5	136 819

Top-Muster	
NN	103 662
NA	76 601
NNN	71 472
NNA	52 574

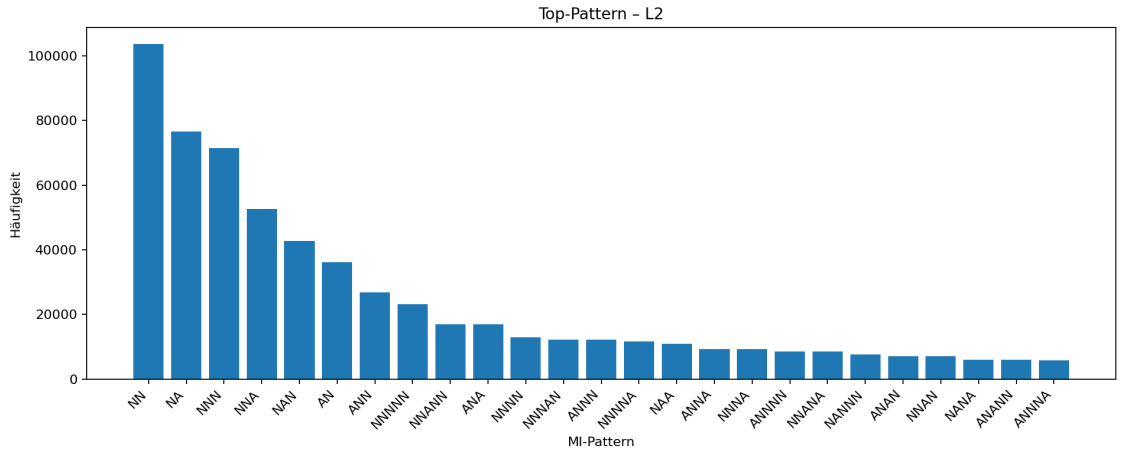

Abbildung 36: Top-Muster - L2: AN(+) Atyp(–) E(+) Longest Matches

Beobachtung. Die Aktivierung des `AN`-Graphen erhöht die Kandidatenzahl um 6 %. Erwartungsgemäß tritt `AN` neu auf und belegt Rang 6, wodurch alle nachfolgenden Muster um eine Position nach unten rücken. Der Effekt der ursprünglichen Designentscheidung wird damit nur teilweise erfasst: Zwar steigt die Zahl der Kandidaten deutlich, das Maß des dadurch erzeugten Rauschens bleibt jedoch ohne qualitative Analyse unbestimmbar.

10.4.3 Lauf 3: Atyp_Syn-Graph aktiviert

Beschreibung. Für diese Arbeit wurden atypische Muster modelliert, die seltene MWEs-Kombinationen zulassen. Diese erwiesen sich als untauglich für die LEMoN-Pipeline, da sie extremes Rauschen erzeugten, und wurden daher für den Hauptlauf deaktiviert.

Konfiguration. `Atyp` = JA; `AN` = NEIN; `<E>` = JA; `Index` = LONGEST MATCHES.

Ergebnisse

Kernzahlen	
Matches (gesamt)	75 710
Dedupliziert	13 289
Lemma-POS	32 951
Jaccard vs. L1	0.119

Längenprofil	
MI=2	1 379
MI=3	14 873
MI=4	13 128
MI=5	46 330

Top-Muster	
NNNNN	7 947
NNANN	5 235
NAN	4 650
NNNAN	4 239

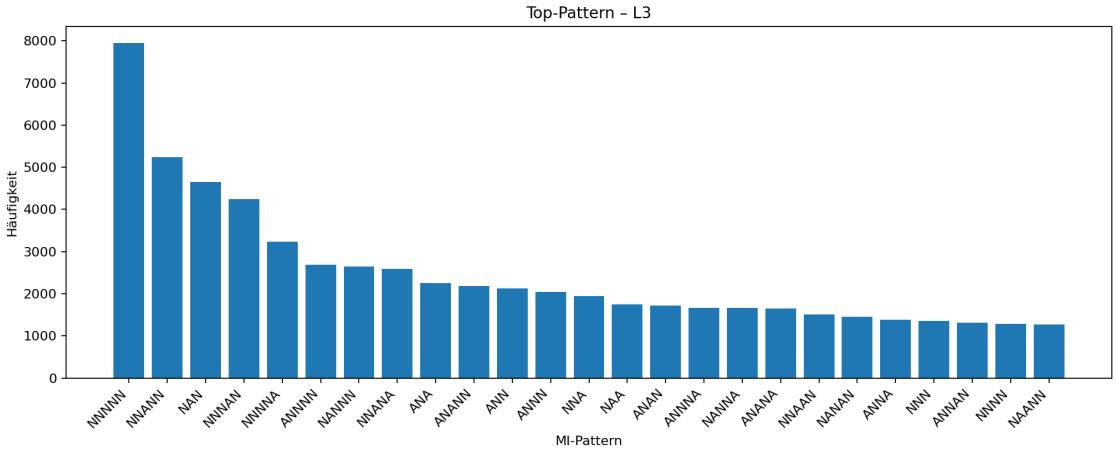

Abbildung 37: Top-Muster - L3: AN(–) Atyp(+) E(+) Longest Matches

Beobachtung. Die Kandidatenzahl sinkt um fast 88 % gegenüber der Baseline. Auffällig ist, dass nun lange Sequenzen mit 5 MI dominieren. Die kurzen Graphen sind zwar noch aktiv, werden jedoch durch die *Longest Matches*-Policy zusammen mit den atypischen Strukturen weitgehend verdrängt. Der Jaccard-Koeffizient von 0,119 bestätigt, dass die Überlappung mit der Baseline nur noch marginal ist.

10.4.4 Lauf 3b: Atyp_Syn-Graph und AN-Graph aktiviert

Beschreibung. Probehalber werden die beiden in L2 (`AN`-Graph) und L3 (atypische Syn) gemessenen Effekte zusammen getestet.

Konfiguration. `Atyp` = JA; `AN` = JA; `<E>` = JA; `Index` = LONGEST MATCHES.

Ergebnisse

Kernzahlen

Matches (gesamt)	76 090
Dedupliziert	13 442
Lemma-POS	33 269
Jaccard vs. L1	0.119

Längenprofil

MI=2	1 760
MI=3	14 873
MI=4	13 128
MI=5	46 329

Top-Muster

NNNNN	7 947
NNANN	5 235
NAN	4 650
NNNAN	4 239

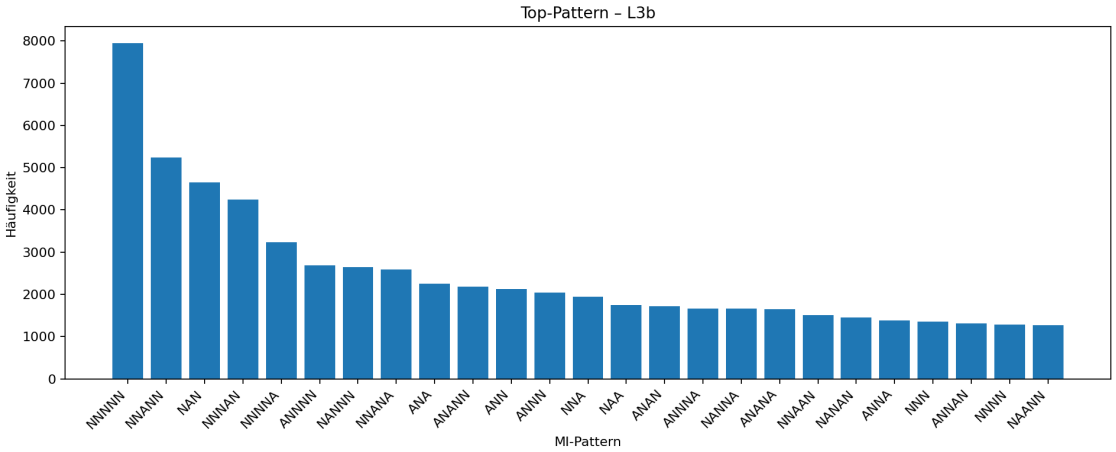

Abbildung 38: Top-Muster - L3b: AN(+) Atyp(+) E(+) Longest Matches

Beobachtung. Der hinzugeschaltete `AN`-Graph hat angesichts der Dominanz der atypischen syntaktischen Graphen kaum einen sichtbaren Effekt: Lediglich wenige hundert Matches bzw. Lemmata sind hinzugekommen. Der Jaccard-Koeffizient ist unverändert, auch die Längenprofile und Topmuster sind kaum beeinflusst. Am interessantesten noch: Es gibt in Lauf L3b eine (!) 5-MI weniger als in Lauf L3.

10.4.5 Lauf 4: <E>-Transitionen deaktiviert

Beschreibung. Testlauf ohne <E>-Transitionen in den btw-Subgraphen, sodass keine „leeren“ Zwischenräume zwischen MI möglich sind.

Konfiguration. Atyp = NEIN; AN = NEIN; <E> = NEIN; Index = LONGEST MATCHES.

Ergebnisse

Kernzahlen	
Matches (gesamt)	340 957
Dedupliziert	108 537
Lemma-POS	155 736
Jaccard vs. L1	0.345

Längenprofil	
MI=2	261 736
MI=3	60 448
MI=4	8 319
MI=5	10 454

Top-Muster	
NN	146 009
NA	115 727
NNN	35 100
NNA	8 710

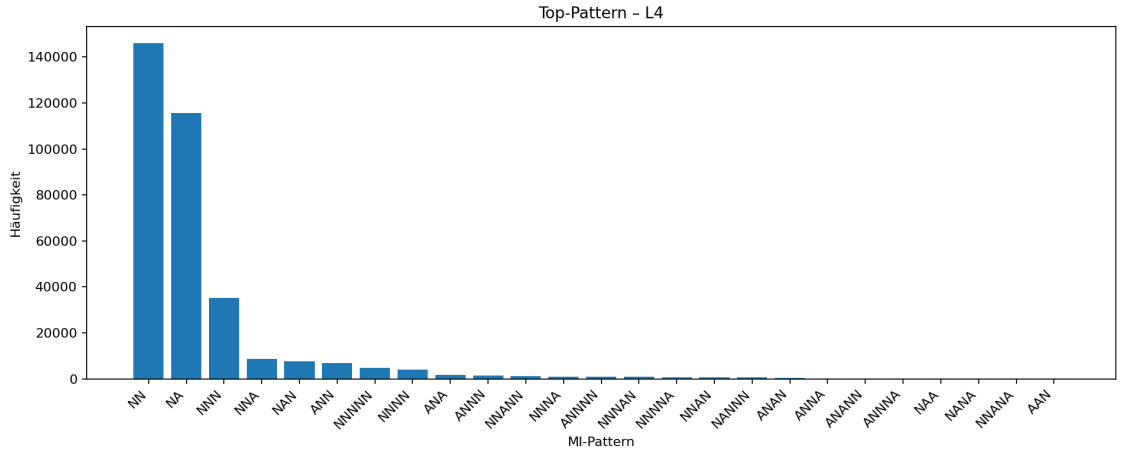

Abbildung 39: Top-Muster - L4: AN(–) Atyp(–) E(–) Longest Matches

Beobachtung. Durch die Deaktivierung der <E>-Transitionen zwischen den MI werden nahezu ausschließlich Kandidaten identifiziert, bei denen Konnektoren wie *de*, *de la*, *de los* etc. zwischen den MI stehen. Entsprechend sinkt vor allem der Anteil längerer Kandidaten sehr stark ab.

10.4.6 Lauf 5: AN, atypische Syn aktiviert, <E>-Transit deaktiviert

Beschreibung. Das Gegenstück zur Baseline: `AN` und atypische Muster sind aktiviert, `<E>`-Transitionen jedoch deaktiviert. Erwartung: Der „Matches-fressende“ Effekt atypischer Strukturen und der durch deaktivierte `<E>`-Transitionen eingeschränkte Suchraum können zusammen ein ungewöhnliches Erkennungsmuster erzeugen.

Konfiguration. `Atyp` = JA; `AN` = JA; `<E>` = NEIN; `Index` = LONGEST MATCHES.

Ergebnisse

Kernzahlen

Matches (gesamt)	11 620
Dedupliziert	3 630
Lemma-POS	5 945
Jaccard vs. L1	0.016

Längenprofil

MI=2	2 290
MI=3	2 627
MI=4	1 539
MI=5	5 110

Top-Muster

NNNNN	2 592
NA	1 068
NNN	1 044
NN	753

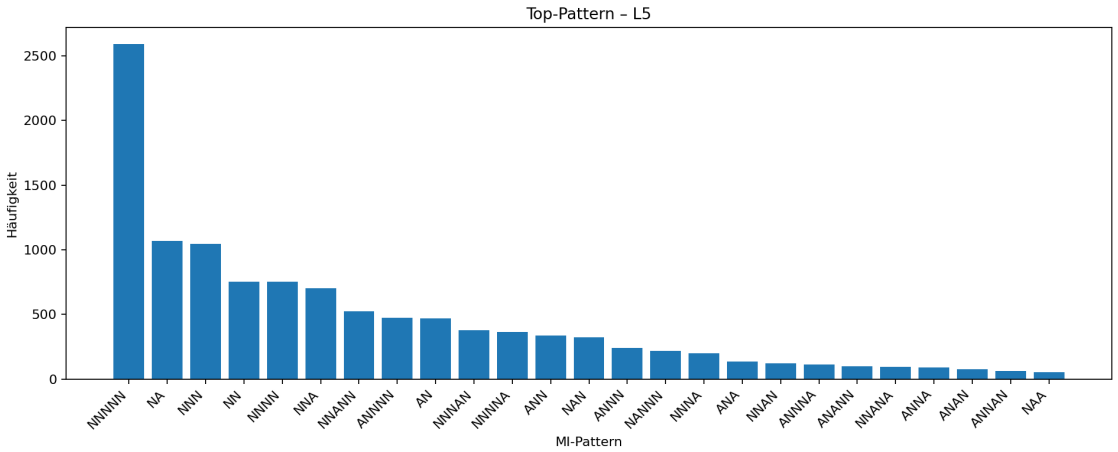

Abbildung 40: Top-Muster - L5: AN(+) Atyp(+) E(–) Longest Matches

Beobachtung. Der Effekt, `AN` und atypische syntaktische Graphen zu aktivieren, bei gleichzeitig deaktivierten `<E>`-Transitionen, reduziert die Matches von 340 957 auf lediglich 11 620. Mit 2 592 Treffern ist überraschenderweise dennoch die 5-MI-Kette `NNNNN` das häufigste Muster. Das sind fast so viele wie die folgenden drei Plätze (`NA`, `NNN`, `NN`) zusammen. Insgesamt ähnelt die Rangkurve der Baseline, allerdings auf sehr niedrigem Niveau.

10.4.7 Lauf 6: Baseline + All Matches

Beschreibung. Der Lauf dient als Diagnostik der *Longest-Match*-Policy. Es werden alle überlappenden Treffer ausgegeben (*All Matches*). Damit wird sichtbar, wie stark verschachtelte Teilsequenzen (z. B. `NN` innerhalb von `NNN`) zum Rohinventar beitragen.

Konfiguration. `Atyp` = NEIN; `AN` = NEIN; `<E>` = JA; `Index` = ALL MATCHES.

Ergebnisse

Kernzahlen	
Matches (gesamt)	1 183 645
Dedupliziert	230 461
Lemma-POS	510 519
Jaccard vs. L1	0.542

Längenprofil	
MI=2	371 482
MI=3	459 887
MI=4	150 693
MI=5	201 583

Top-Muster	
NN	222 208
NNN	152 226
NA	149 274
NNA	96 178

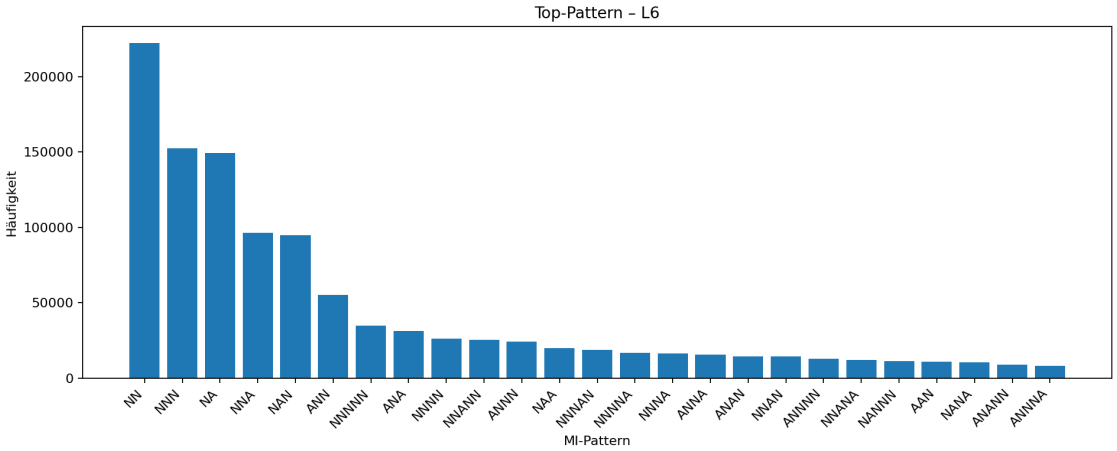

Abbildung 41: Top-Muster - L6: AN(–) Atyp(–) E(+) *All Matches*

Beobachtung. Bei Berücksichtigung aller Matches, anders gesagt, ohne Festlegung auf *Longest-Matches*, verdoppelt sich die Zahl der Matches und beinahe aller anderen Werte ungefähr - die einzige Ausnahme bildet der Wert für 5 MI-Matches. Es ist davon auszugehen, dass viele mehrgliedrige Kandidaten jetzt in ihre Elemente aufgespalten vorliegen. Das belegt den *Longest-Match*-Effekt: robustere, aber kleinere Inventare (vgl. L1), was sich für den konservativen Hauptlauf anbietet.

10.5 Vergleich der Ablationsläufe

Dieses Kapitel integriert die Ergebnisse des Hauptlaufs sowie der Ablationsläufe (L1-L6). Während der Hauptlauf die Referenzkonfiguration für die Gesamtauswertung darstellt, ist L1 als Baseline innerhalb der Ablationsstudie angelegt und entspricht in den Parametern dem Hauptlauf. Vergleiche zum Hauptlauf sind als Trendaussagen zu lesen. Numerische Gegenüberstellungen erfolgen innerhalb der Ablation relativ zu L1. Ziel ist es, die getroffenen Designentscheidungen in der LEMoN-Pipeline transparent zu bewerten und die Robustheit der Referenzkonfiguration einzuordnen.

Nachdem die einzelnen Läufe vorgestellt und bewertet wurden, ist ein direkter Vergleich besonders aufschlussreich. Die folgende Abbildung stellt die Kernzahlen der Läufe L1 bis L6 graphisch dar. Dadurch lässt sich unmittelbar erkennen, welchen Effekt die jeweiligen Einstellungen auf die Gesamtzahl der Matches, die dedupliziert-ten Matches (d. h. wirklich „neue“ Treffer) sowie die Lemma-POS-Paare (d. h. die lexikalisierungsrelevanten Einheiten) haben.

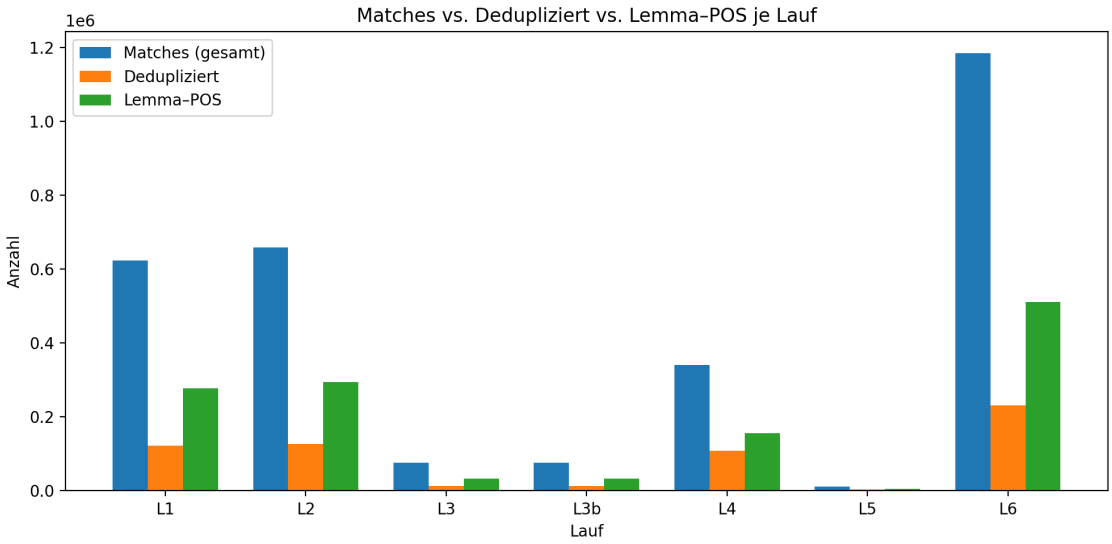

Abbildung 42: Vergleich der Kernzahlen je Lauf

Die Abbildung zeigt besonders den Effekt der Index-Policy All Matches in Lauf L6. Der Balken für die Gesamtmatches sticht gegenüber L1 bis L5 deutlich hervor. Gleichzeitig wachsen auch die deduplizierten Matches und die Lemma-POS-Paare substantiell, wenn auch weniger ausgeprägt: gegenüber L1 etwa 230 461 statt 122 035 deduplizierte Treffer und 510 519 statt 276 918 Lemma-POS-Paare. All Matches erhöht damit die Rohmenge am stärksten und vergrößert zugleich das inventarrele-

vante Material deutlich. Auffällig ist zudem der Effekt der atypischen syntaktischen Graphen: L3, L3b und L5 sind unmittelbar an der stark reduzierten Trefferzahl zu erkennen. Das Deaktivieren der `<E>`-Transitionen in L4 wiederum beschneidet Matches und Lemmata in ähnlicher Größenordnung, wobei die Zahl der deduplizierten Matches bemerkenswert nah an der Baseline (L1) bleibt.

Die nächste Abbildung zeigt die Jaccard-Koeffizienten der verschiedenen Läufe im Überblick:

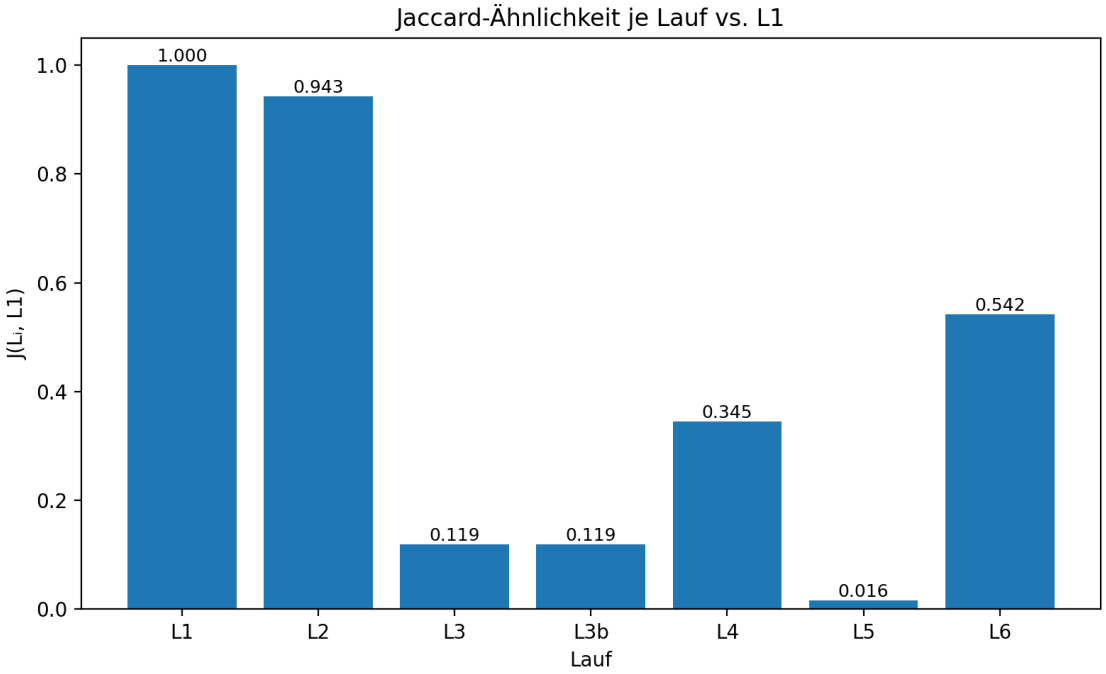

Abbildung 43: Jaccard-Ähnlichkeit der Lemma-POS-Paare je Lauf

Wie erwartet liegt L2 sehr nah an der Baseline. Der Unterschied, ob `AN` aktiviert ist oder nicht, bleibt verhältnismäßig gering. Dass L5 als „Gegenstück zur Baseline“ bezeichnet wurde, zeigt sich auch unmittelbar an der großen Distanz: Der Jaccard-Wert beträgt nur 0,016. Überraschend ist hingegen, dass die Anwendung der *All Matches*-Index-Policy den Abstand kaum vergrößert: L6 ist nach L2 immer noch der zweitnächste Lauf zur Baseline. Damit ist L6 trotz deutlich größerer Rohmenge inhaltlich näher an L1 als alle Konfigurationen mit `syn_atyp`. Die Abweichung entsteht vor allem durch zusätzliche, überlappende Varianten.

In Summe bestätigt die Ablation die Hauptlauf-Konfiguration als robusten Kompromiss. `AN` erweitert moderat, `syn_atyp` verzerrt stark, `<E>` ist für längere Ketten zentral und *All Matches* liefert die größte Abdeckung bei deutlich höherem Auswer-

tungsaufwand.

Nach der Darstellung der Kernzahlen und der Ähnlichkeit zur Baseline folgt zuletzt der Blick in die einzelnen Läufe hinein. Abbildung 44 zeigt eine Heatmap, in der für alle syntaktischen Kombinationen (vertikal) die Zahl der Matches pro Lauf (horizontal) visualisiert ist. Die Farbskala (rechts) ist logarithmisch von 10^{-9} bis 10^{5} gestreckt und erlaubt damit auch den Vergleich sehr ungleicher Häufigkeiten: Dunkelblau steht für sehr häufige, Hellblau für seltene und Weiß für nicht vorkommende Muster. Zur besseren Lesbarkeit sind nur Zellen mit mehr als 100 Treffern beschriftet. Auffällig ist, dass in L4 und L5 mehrere längere Kandidaten, die in anderen Läufen identifiziert wurden, vollständig entfallen.

Die Effekte der einzelnen Ablationsläufe lassen sich klar voneinander unterscheiden: In L1 und L2 dominieren nominale Muster, insbesondere `NN`, `NNN`, `NNNN` und `NNNNN`. In L2 kommt natürlich neu das Muster `AN` hinzu. In L3 und L3b brechen die 2-MI- und 3-MI-Matches nahezu vollständig ein. Längere Muster bleiben zwar erhalten, sind aber im Vergleich stark reduziert. In L4 und L5 zeigt die Deaktivierung der `<E>`-Transitionen einen Effekt über alle MI-Längen hinweg – mit Ausnahme der 2-MI-Matches, die in L4 deutlich häufiger als in der Baseline auftreten. In L5 verschwindet auch dieser Effekt, da das Hinzufügen der atypischen syntaktischen Graphen die Erkennung insgesamt zum Erliegen bringt. L6 mit der *All Matches*-Index-Policy weist demgegenüber auf allen MI-Längen hohe Treffermengen auf.

Neben den von vornherein ausgeschlossenen nicht-nominalen Graphen (z,B. `AA`) sind auch die Muster `AAAAN`, `NAAAA`, `NANAA`, `NNAAA` und `NANN` trefferlos und erscheinen daher weiß in der Heatmap. Die Graphen `AAAAN`, `NAAAA` und `NNAAA` wurden in frühen Iterationen wegen starker adjektivischer Prägung nicht implementiert. `NANAA` wurde ebenfalls nicht umgesetzt, hier ist von einem Versäumnis auszugehen. `NANN` ist zwar modelliert und implementiert, wurde jedoch durch einen Fehler im Transducer-Output irrtümlich als `ANNN` ausgegeben. Um die Auswirkungen transparent zu prüfen, habe ich einen internen Lauf (L7) mit korrigiertem Output und Baseline-Konfiguration durchgeführt. Für das getestete Teilkorpus ergab sich eine Trefferzahl von null. Damit sind die Ablationsläufe von diesem Fehler nicht betroffen. Im Hauptlauf wurden diese Muster angelegt und korrigiert. Es liegt Vollabdeckung vor, und alle modellierten 55 Muster wurden beobachtet.

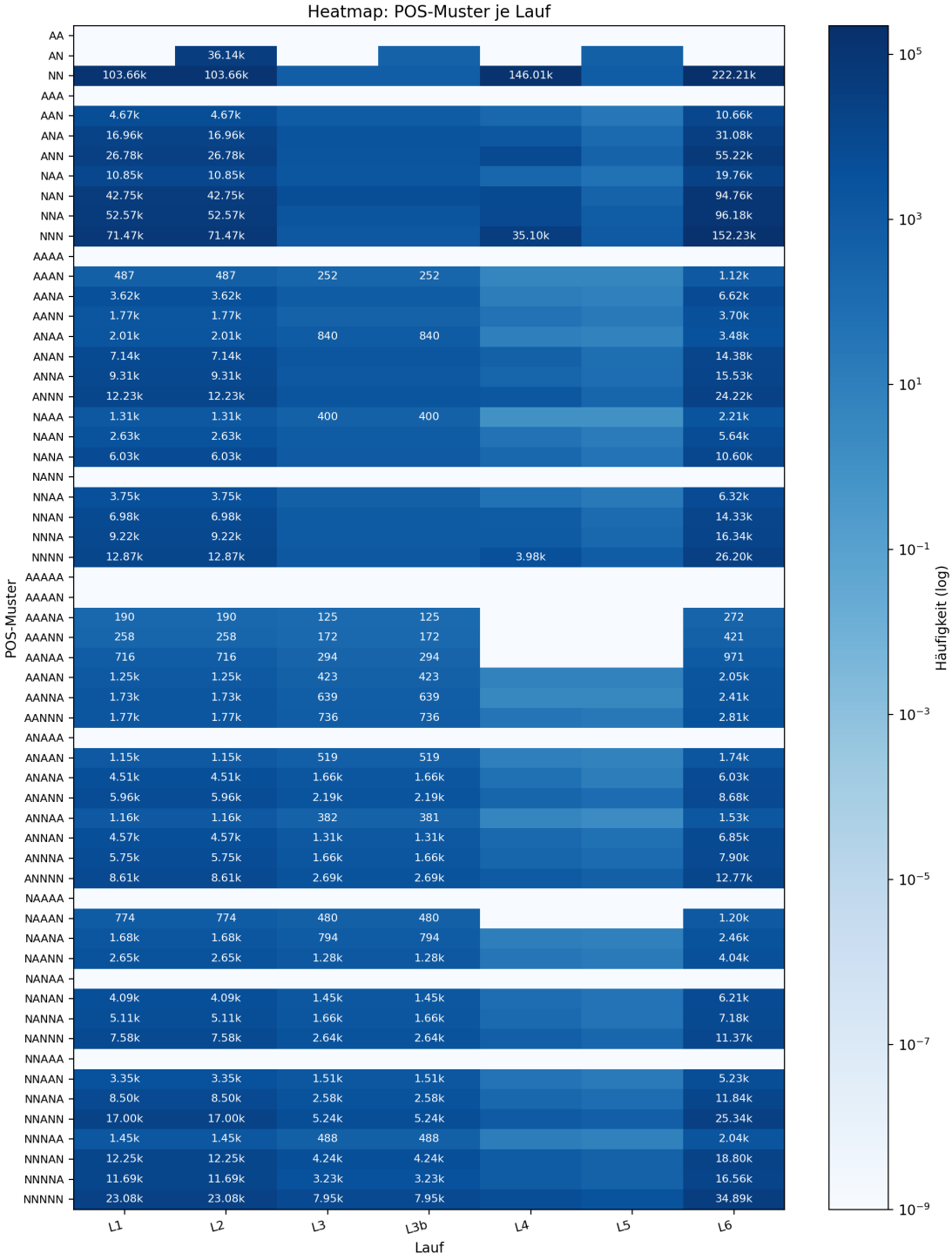

Abbildung 44: Heatmap der POS-Muster über die Ablationsläufe (L1-L6)

10.6 Bewertung der Designentscheidungen

Leitfragen

1) Erhöht `AN` die Trefferqualität oder vor allem das Rauschen, und wie groß ist der Nettogewinn nach Deduplikation auf Lemma-POS-Ebene?

2) Sind `<E>`-Transitionen im Hinblick auf Abdeckung längerer MI-Ketten unverzichtbar, und sind die dadurch induzierten Fehler vertretbar?

3) Lässt sich der Verzicht auf `syn_atyp` rechtfertigen, gegeben deren Einfluss auf Verteilung und Vergleichbarkeit der Muster?

4) Welche Unterschiede entstehen durch `All Matches` gegenüber `Longest Matches` in Bezug auf Verschachtelungen, Arbeitsaufwand und Informationsgewinn?

5) Wie robust ist die Baseline-Konfiguration gegenüber den Alternativen insgesamt zu bewerten?

10.6.1 Der Einfluss von `AN`

Die Deaktivierung des `AN`-Graphen wirkt kontraintuitiv, denn adjektivische nominale MWEs sind keineswegs exotisch. Im Bootstrapping zeigte sich jedoch, dass `AN` viel Rauschen erzeugt. L2 quantifiziert das: Mit aktiviertem `AN` steigt die 2-MI-Klasse von 180 267 auf 216 405 (+36 138 ≈ +6,0 %). Insgesamt wächst die Matchzahl gegenüber L1 um +36 133 (+5,8 %). Nach Deduplikation bleiben davon +4 344 (+3,6 %) übrig, auf Lemma-POS-Ebene sind es +16 746 neue Paare (+6,0 %). `AN` selbst steigt – erwartbar – neu in die Top-Muster auf (Rang 6). Der starke Rückschnitt durch Deduplikation zeigt, dass `AN` vor allem viele identische oder sehr ähnliche Varianten findet. Zur inhaltlichen Qualität sagt das nichts – die bräuchte eine gesonderte, qualitative Sichtung. Der hohe Jaccard-Wert (0,943) bestätigt zugleich: L2 bleibt inhaltlich sehr nah an der Baseline. `AN` erweitert das Inventar moderat, aber nicht disruptiv, und bleibt im Verhältnis zu `NN`+`NA` (zusammen 180 000+) deutlich kleiner.

Bei der Analyse der Kernzahlen fällt zudem auf, dass es auch geringe Bewegungen außerhalb der Erwartung gibt: So verliert das 5-MI-Längenprofil fünf Matches. Zeitgleich weisen die Top-Muster `NN` und `NA` jeweils zwei Treffer weniger auf. Ob es sich

hierbei um statistische Artefakte handelt – etwa bedingt durch den Paumier-Search-Algorithmus (laut Unitex-GUI im Vergleich zur Automaton Intersection „quicker“, aber „less precise“, jedoch nicht vollständig dokumentiert, vgl. [S. Paumier 2016]) – oder um eine Eigenheit der *Longest Matches*-Index-Policy, bleibt offen. Letztere könnte nicht global, sondern nur graph-intern die jeweils längsten Matches bevorzugen und dadurch beim Hinzutreten neuer Kandidaten Verschiebungen erzeugen.

Bewertung: `AN` nicht zu aktivieren, würde ich heute anders entscheiden. Zwar ist zu erwarten, dass die zusätzlichen Treffer überproportional Rauschen enthalten, doch hat die damalige Wahl in einem *Recall-orientierten* Ansatz eine unnötige und zudem basale Leerstelle in der Erkennung geschaffen.

10.6.2 Der Verzicht auf atypische syntaktische Muster

Die Funktion der Graphen für asyntaktische Muster ist bereits mehrfach beschrieben worden. Ihr Effekt auf die Erkennung ist erheblich und auf den ersten Blick unintuitiv: Das fast völlige Versiegen der typischen Muster dürfte eine Folge der *Longest Matches*-Policy sein, wirft aber Fragen für weiterführende Studien auf. Den Einfluss habe ich in drei Ablationsläufen untersucht: isoliert (L3), in Kombination mit `AN` als erweiterte Baseline (L3b) sowie im „Gegenstück“-Lauf L5. In allen Fällen war der Effekt der Graphenfamilie deutlich erkennbar – jedoch primär auf die Zahl der erkannten Muster, nicht auf deren Qualität (die modellierten Strukturen waren deutlich diverser, die resultierende Datenbasis zugleich reduziert).

Bewertung: Die Deaktivierung war richtig. Der Einfluss auf die typischen Strukturen ist zu stark, um beide Ansätze gemeinsam im Hauptlauf zu verfolgen. Sinnvoller wäre eine separate Anwendung parallel zum Hauptlauf gewesen, verbunden mit einer gesonderten qualitativen Analyse der Ergebnisse. Hier wäre dann auch der Einsatz von *All Matches* zu prüfen.

10.6.3 Deaktivierung von <E>-Transitionen

Die `<E>`-Transitionen verbinden zwei MI ohne expliziten Konnektor. Das ermöglicht die Erkennung vieler nominaler MWEs, erhöht jedoch auch das Risiko, nicht zusammengehörige Wörter zu überspannen. Lauf L4 prüft den Effekt gezielt. Die Subgraphen `btwNN`, `btwNA`, `btwAN` und `btwAA` wurden so geändert, dass sie keine `<E>`-Transitionen mehr enthalten. Dadurch entsteht eine starke Einschränkung. Alle MWEs, deren MI nicht über Elemente wie *de la* oder *de los* verbunden sind, werden nicht mehr erkannt. Entsprechend sinken die Matchzahlen deutlich, längere Ketten

verschwinden nahezu vollständig, der Jaccard-Koeffizient fällt auf 0,345. Das unterstreicht die zentrale Rolle präpositionaler Verbindungen für spanische nominale MWEs.

Bewertung: `<E>`-Transitionen sind für einen erheblichen Teil der Matches wichtig, besonders bei längeren MI-Ketten. Ihre Deaktivierung macht den Einfluss sichtbar, erlaubt aber keine genaue Aussage zur Auswirkung auf die Präzision. Dafür wäre eine zusätzliche qualitative Bewertung erforderlich.

10.6.4 Die Wahl der Index-Policy

Die in Unitex voreingestellte Standard-Policy *Longest Matches* wurde übernommen, weil es methodisch plausibel erschien, zunächst geschachtelte MWEs zu identifizieren und ihre Bestandteile bei Bedarf nachgelagert über das neu erzeugte `LEMoN.dic`-Lexikon zu erfassen. *All Matches* hätte zwar beides zugleich geliefert, schien mir im Bootstrapping jedoch eher störendes Rauschen zu erzeugen.

Der Effekt zeigt sich in Lauf L6 deutlich: In allen drei Metriken verdoppeln sich die Kernzahlen in etwa, ebenso die meisten Muster – lediglich bei den Kandidaten mit 5 MI fällt der Anstieg etwas geringer aus (von rund 136 824 auf 201 583). Entsprechend sinkt der Jaccard-Koeffizient auf etwa die Hälfte. Die manuelle Sichtung wird dadurch erheblich aufwändiger, da identische oder verschachtelte Kandidaten mehrfach auftreten.

Bewertung: Die Entscheidung ist ambivalent: Für die Entwicklung und die Baseline-Konfiguration liefert *Longest Matches* robustere, leichter interpretierbare Daten. Für eine recall-getriebene Auswertung (insbesondere im Hauptlauf) wäre aus heutiger Sicht jedoch evtl. *All Matches* vorzuziehen - auch wenn dies den Aufwand der manuellen Analyse deutlich erhöht. Es ist außerdem nicht von einem tatsächlichen Anwachsen in dieser Größenordnung auszugehen, stattdessen wird ein erheblicher Teil durch die Deduplizierung entfallen.

10.6.5 Gesamtbewertung der Konfiguration für den Hauptlauf

Die Ablationsstudie war ein geeignetes Mittel, um verschiedene Designentscheidungen in den LEMoN-Grammatiken auf den Prüfstand zu stellen. In der Gesamtschau zeigt sich die gewählte Hauptlaufkonfiguration als praktikabler, jedoch keineswegs idealer Kompromiss. Jeder der getesteten Parameter hätte die Ergebnisse auf eine bestimmte Weise verschoben – jedoch stets um den Preis neuer Probleme.

So hätte die Aktivierung des `AN`-Graphen zwar einen Zuwachs von rund sechs Prozent gebracht, die damit verbundenen Treffer sind jedoch größtenteils reguläre und wenig informative Kombinationen – in deutlich stärkerem Maße als bei den anderen 2-MI-Kandidaten (dies war auch die ursprüngliche Motivation für die Deaktivierung). Die Integration der atypischen syntaktischen Graphen wiederum hätte nicht nur die Treffermenge um fast 88 % reduziert, sondern zugleich die typischen Strukturen so stark verfremdet, dass eine Vergleichbarkeit der Ergebnisse kaum noch möglich gewesen wäre. Die Deaktivierung der `<E>`-Transitionen hätte die Präzision zwar deutlich gesteigert, jedoch um einen hohen Preis: Eine ganze Klasse relevanter Kandidaten wäre ausgeschlossen worden - mit besonders drastischen Einbußen bei längeren MI-Ketten. Schließlich hätte die Index-Policy *All Matches* die Abdeckung zwar erheblich vergrößert, jedoch um den Preis einer Verdopplung der Datenbasis. Dies hätte die manuelle Analyse praktisch unbewältigbar gemacht, da identische oder verschachtelte MWEs mehrfach auftreten und dadurch sowohl die Auswertung als auch das gezielte Design von Grammatiken erheblich erschweren.

Vor diesem Hintergrund erscheint die Baseline-Konfiguration als die Wahl mit den wenigsten Einschränkungen bzw. dem stabilsten Kompromiss: Sie vermeidet einerseits extreme Verzerrungen und wahrt andererseits eine akzeptable Balance zwischen Recall und Präzision, wodurch eine methodisch kontrollierbare Auswertung ermöglicht wird. Gleichwohl handelt es sich nicht um eine optimal justierte Lösung, sondern um eine pragmatische Konfiguration innerhalb der gegebenen Rahmenbedingungen. Positiv gewendet lässt sie sich als robuste Konfiguration bewerten: Die Designentscheidungen sind nachvollziehbar und entfalten im Wesentlichen den intendierten Effekt.

Weitere Studien sind notwendig, um diese und weitere Stellschrauben in der LEMoN-Pipeline systematisch und unter kontrollierten Bedingungen zu variieren und ihre Effekte differenzierter zu quantifizieren.

10.7 Methodische Einschränkungen

Mit Blick auf diese Arbeit möchte ich auch ein paar methodische Einschränkungen hervorheben, in deren Licht sowohl die LEMoN-Pipeline, die LEMoN-Grammatiken, das Resultat `LEMoN.dic` sowie die Ablationsläufe und ihre Analyse stehen.

10.7.1 Recall-Orientierung

Wie mehrfach beschrieben basiert `LEMoN` auf einer *Recall-orientierten* Strategie: Ziel ist eine möglichst breite Erfassung potenzieller Kandidaten, auch auf Kosten der Präzision. Zwar erlauben lokale Grammatiken prinzipiell das Gegenteil, denn durch extrem präzise Kontexte lassen sich mit hoher Sicherheit nur tatsächlich vorkommende MWEs identifizieren. Dieser Ansatz hätte jedoch das zentrale Problem der MWE-Lexikographie ignoriert: Die Zahl existierender, bislang nicht erfasster MWEs ist unbekannt, aber mutmaßlich extrem hoch, und die meisten treten nur sehr selten auf. Eng gefasste Kontexte würden daher in erster Linie bereits bekannte MWEs reproduzieren, anstatt neue Kandidaten sichtbar zu machen.

Vor diesem Hintergrund wurde bewusst eine breitere Abdeckung gewählt. Das erschwert allerdings die Evaluation: Ein kuratiertes Referenzkorpus zur direkten Messung des Recalls liegt nicht vor und hätte den Rahmen dieser Arbeit überstiegen. Die Evaluation fokussiert daher primär auf Präzision, die stichprobenbasiert erhoben wurde. Rückschlüsse auf den Recall lassen sich nur indirekt aus den Ablationsläufen und der Größe des Kandidateninventars ableiten.

Diese Einschränkung verweist zugleich auf ein grundlegendes methodisches Dilemma: Recall lässt sich nur auf Kosten der Präzision steigern. Die vergleichsweise niedrige Präzision, insbesondere in aggressiven Konfigurationen, bestätigt die Recall-orientierte Strategie: Auch schwächere Kandidaten werden erfasst, um eine maximale Abdeckung zu gewährleisten. Für eine finale Ressourcenerstellung wäre dies unbefriedigend, für die Evaluation der Methodik jedoch erwartbar und methodisch folgerichtig.

10.7.2 Korpuswahl

Die Wahl des *EuroParl*-Korpus hatte unmittelbare Konsequenzen für die Ergebnisse und ihre Übertragbarkeit. *EuroParl* ist durch seine formale Registerprägung stark fachsprachlich ausgerichtet und begünstigt Strukturen mit mehr als 4 MI. Hinzu kommt der Einfluss von *translationese*, da ein erheblicher Teil des Materials auf Übersetzungen basiert. Ursprünglich war vorgesehen, zunächst ein kleines Zeitungskorpus (El Mundo) für den Aufbau der Grammatiken und anschließend ein umfangreicheres Zeitungskorpus (El País) für die eigentliche Evaluation zu nutzen. Dieses Vorhaben scheiterte jedoch an der Größe und Komplexität des El País-Korpus. Dadurch ergab sich ein Wechsel des Registers, der die Vergleichbarkeit mit

Alltagssprache einschränkt, während die Wahl von *EuroParl* pragmatisch durch seine Verfügbarkeit, Standardisierung und Dokumentation begründet war.

10.7.3 Wörterbuchqualität

Einige auffällige Ausreißer (z. B. extrem viele Labels pro Lemma) sind auf Fehler in den DELA-Ressourcen oder auf Koordinationsartefakte zurückzuführen. Solche Probleme wären prinzipiell durch manuelle Kuration (z. B. engere Graphen oder gezielte Korrekturen im Lexikon) zu beheben, liegen jedoch im aktuellen Output vor. Hinzu kommt, dass die OOV-Wörterbücher in einem aufwendigen, überwiegend nebenberuflichen Verfahren erstellt wurden und daher teilweise inkonsistent annotiert sind. Dies hat jedoch nur Auswirkungen auf das konkret hier erzeugte `LEMoN.dic` und seine Präzision, weniger auf das dahinterstehende Prinzip: In Verbindung mit der überwiegend generischen Graphenarchitektur, die im Wesentlichen auf Sequenzen wie `DET`, `AN` oder `NNA` sowie einigen wenigen spezifisch modellierten Konstruktionen (z. B. *a favor de*) beruht, ergibt sich eine vergleichsweise robuste Grammatikstruktur. Daraus folgt, dass die Pipeline nicht an die aktuell verwendeten, teilweise unsauber annotierten Wörterbücher gebunden ist: Mit besser kuratierten Lexika und einem passenden Korpus ließen sich unmittelbar qualitativ hochwertigere Ergebnisse erzielen, ohne dass ein grundlegendes Redesign der Graphen erforderlich wäre.

10.7.4 Weitere

Weitere methodische Einschränkungen betreffen die folgenden - hier nicht weiter vertieften - Aspekte:

- Suchalgorithmen: Unterschiede zwischen Paumier Search und Automaton Intersection mit potenziellen Auswirkungen auf die Treffer (nicht vollständig dokumentiertes Feature von Unitex).
- Umfang der Ablation: begrenzte Größe der Untersuchung (Teilkorpus, Stichprobe von 50 Kandidaten).
- Analysefaktoren: bewusster Verzicht auf semantische Klassen als zusätzlichen Faktor.

Teil IV

Abschluss

11 Fazit

Lo mejor será escoger el camino de Galta, recorrerlo de nuevo (inventarlo a medida que lo recorro) y sin darme cuenta, casi insensiblemente, ir hasta el fin -sin preocuparme por saber qué quiere decir „ir hasta el fin“ ni qué es lo que yo he querido decir al escribir esa frase.

Octavio Paz

11.1 Zusammenfassung

Diese Arbeit verfolgte zwei komplementäre Perspektiven: zum einen die Gewinnung eines Wörterbuchs nominaler Mehrworteinheiten für das Spanische, `LEMoN.dic`, zum anderen die experimentelle Untersuchung der LEMoN-Pipeline selbst. Entsprechend richtet sich der Blick im Folgenden sowohl auf die Ressource `LEMoN.dic`, die das Ergebnis des Hauptlaufs der Pipeline auf das *EuroParl*-Korpus repräsentiert, als auch auf die Robustheit und Grenzen der eingesetzten Methodik, die durch die Ablationsläufe sichtbar wurden.

Die in dieser Arbeit vorgestellten Prinzipien und Ressourcen, denen ich den Namen LEMoN gegeben habe, haben das Ziel, pragmatisch und ressourcenbewusst ein Lexikon spanischer nominaler Mehrworteinheiten zu erstellen. Dieses Vorgehen spiegelte nicht nur meine persönliche Arbeitssituation wider - mehrfach änderten sich während der Bearbeitung die äußeren (Lebens-)Umstände -, sondern entspricht auch dem gewählten Ansatz. Die Schule von Maurice Gross, deren Betonung manueller Lexikonarbeit sich als kräftezehrend, zugleich aber als intellektuell tragfähig erwiesen hat, bietet auch nach der langen Bearbeitungszeit ein faszinierendes Fundament. Die Überzeugung, dass die Idee, Sprache als Kontinuum von Lexikon und Grammatik zu betrachten, auch in Zeiten KI-basierter Modelle zentral und unterschätzt bleibt, hat mich durch schwierige Phasen getragen.

Methodisch stand zum einen die Entwicklung einer Pipeline im Mittelpunkt, die ausgehend von Korpora und Wörterbüchern über lokale Grammatiken ein Kandidateninventar spanischer nominaler Mehrworteinheiten erzeugt. Zum anderen erlaubte eine Ablationsstudie, zentrale Designentscheidungen systematisch zu überprüfen. Die Ergebnisse zeigen, dass LEMoN robuste Resultate liefert, zugleich aber zahlreiche Stellschrauben und Optimierungsmöglichkeiten offenlässt.

Das resultierende Kandidatenlexikon `lemon.dic` enthält in erheblichem Umfang reguläre Nominalphrasen und - mit angemessener Konfidenz - auch genuine Mehrworteinheiten. Diese Inklusivität ist mit dem Recall-orientierten Ansatz in Kauf genommen worden, kann aber manuell kuratiert werden. Die LEMoN-Grammatiken haben sich als tragfähig, aber verbesserungsfähig erwiesen: Ihre generische Architektur erlaubt es, unmittelbar von besser kuratierten Wörterbüchern oder alternativen Korpora zu profitieren. Mit einem geeigneten Referenzkorpus ließe sich zudem der Recall direkt messen - etwas, das in dieser Arbeit aus Ressourcengründen nicht möglich war.

11.2 Ausblick

Für die Zukunft ergibt sich ein klarer Arbeitsplan: `LEMoN.dic` müsste zunächst manuell kuratiert werden, um von einem Kandidatenlexikon zu einem eigentlichen Lexikon komplexer Einheiten zu werden. Schon das einfache Entfernen objektiv falscher Einträge würde die Ressource für viele computerlinguistische Anwendungen nutzbar machen. Eine umfassendere Bearbeitung - mit Entfernung bloßer Nominalphrasen, Korrektur fehlerhafter POS-Annotationen und Erweiterung durch spezifischere Grammatiken - könnte LEMoN schließlich zu einem hochwertigen, nachhaltigen Beitrag zur lexikalischen Ressourcensammlung für das Spanische machen.

Darüber hinaus eröffnen sich mehrere Perspektiven:

- **Integration in bestehende Ressourcen.** Eine Anbindung von `LEMoN.dic` an Projekte der MWE-Community könnte die langfristige Nutzung sichern und zugleich eine Grundlage für weiterführende Untersuchungen und Erweiterungen schaffen.
- **Übertragung auf weitere Sprachen.** Die generische Architektur der LEMoN-Grammatiken erlaubt eine einfache Adaption auf andere romanische Sprachen, und darüber hinaus. Das parallele *EuroParl*-Korpus eröffnet zudem die Möglichkeit, ein mehrsprachiges Netzwerk nominaler MWE-Ressourcen aufzubauen.
- **Kombination mit KI-Methoden.** Besonders vielversprechend erscheint die Verbindung regelbasierter Extraktion mit neuronalen Sprachmodellen. Moderne LLMs arbeiten zwar erfolgreich mit großen Textmengen, doch sie behandeln

Mehrworteinheiten oft nur oberflächlich als Abfolgen von Token. LEMoN könnte hier als lexikalisches Rückgrat dienen: nicht als „Rohmaterial fürs Training“, sondern als strukturierte Zusatzinformation, etwa für *syntax- und lexikonbewusste Transformer-Modelle*. Ebenso bietet sich LEMoN für die Evaluation von LLMs an, indem überprüft wird, inwieweit komplexe nominale Einheiten korrekt erkannt, übersetzt oder paraphrasiert werden.

Die Arbeit an einem Lexikon ist nie abgeschlossen. LEMoN ist bewusst als offenes, erweiterbares Verfahren angelegt, denn Sprache verändert sich, ebenso die Welt, in der sie verwendet wird, und auch die technischen Mittel, mit denen wir sie erschließen. Umso bemerkenswerter ist es, dass bislang noch kein wirklich vollständiges Inventar ihrer basalen Einheiten existiert. LEMoN ist in diesem Sinne nur ein kleiner Schritt im Verhältnis zur Größe der Aufgabe - aber es ist ein Schritt, der gegangen ist, und damit ein Fundament, auf dem aufgebaut werden kann.

Gerade weil Mehrworteinheiten für KI-Systeme trotz beachtlicher Fortschritte nach wie vor eine Grenze darstellen, bestätigt sich - mit einem Augenzwinkern - Wittgensteins Satz: „Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.“ Solche Grenzen sind jedoch verschiebbar: durch systematische Erschließung, lexikographische Sorgfalt und die Verbindung traditioneller Methoden mit neuen Technologien. Ob und in welchem Maße diese Arbeit dazu einen Beitrag leisten kann, muss seinerseits Gegenstand weiterer Untersuchungen bleiben.

> *Wie wenn tagelang feine, dichte Flocken vom Himmel nieder fallen, bald die ganze Gegend in unermeßlichem Schnee zugedeckt liegt, werde ich von der Masse aus allen Ecken und Ritzen auf mich andringender Wörter gleichsam eingeschneit.*
>
> Jacob Grimm

Teil V
Anhänge

A Tabellen zur Lexikographierung von OOV

Hinweise und Legende Die Tabellen dokumentieren Rauschquellen, Heuristiken und typische Fehlklassifikationen. Abkürzungen: N = Nomen, A = Adjektiv, V = Verb, Genus/Numerus: m = maskulin, f = feminin, s = Singular, p = Plural.

Tabelle 20: Zeichen mit erhöhter Artefaktwahrscheinlichkeit als Kandidaten für das Entfernen vor der Lexikographierung (Gremlins)

Zeichen	Unicode	Beschreibung
©	U+00A9	Copyrightzeichen
®	U+00AE	Registriertes Markenzeichen
™	U+2122	Warenzeichen
§	U+00A7	Paragraphenzeichen
¶	U+00B6	Absatzzeichen
•	U+2022	Aufzählungszeichen
–	U+2013	Halbgeviertstrich
—	U+2014	Geviertstrich
…	U+2026	Auslassungspunkte
°	U+00B0	Gradzeichen
±	U+00B1	Plusminuszeichen
→	U+2192	Pfeil nach rechts
←	U+2190	Pfeil nach links
↑	U+2191	Pfeil nach oben
↓	U+2193	Pfeil nach unten

Tabelle 21: Indikatoren für Fremdsprachenkontext: Funktionswörter aus anderen Sprachen (Auswahl)

Form	Sprache	Wortart/Funktion
of	Englisch	Determinativ
the	Englisch	Determinativ
to	Englisch	Präposition
di	Italienisch	Präposition
do	Portugiesisch	Präposition
da	Italienisch	Präposition
der	Deutsch	Artikel
die	Deutsch	Artikel
das	Deutsch	Artikel
du	Deutsch	Pronomen
il	Französisch/Italienisch	Artikel/Pronomen
ai	Französisch	Hilfsverbform
be	Englisch	Verb Grundform
am	Englisch	Hilfsverbform
je	Französisch	Pronomen
my	Englisch	Determinativ
it	Englisch	Pronomen

Tabelle 22: Diakritische Zeichen, die im Spanischen nicht vorkommen (Auswahl)

Diakritika	Beispielwort	Sprache	Name
Ăă	măcel	Rumänisch	Breve
Șș	școală	Rumänisch	Unterkomma
Çç	garçon	Französisch	Cedille
Èè	très	Französisch	Gravis
ß	Straße	Deutsch	Eszett
Łł	łódź	Polnisch	L mit Schrägstrich
Ąą	ąsz	Polnisch	Ogonek
Ěě	město	Tschechisch	Hatschek
Đð/đ	đak	Kroatisch	D mit Querstrich
Ćć	Ćevapčići	Kroatisch	Akut
ı	Nişantaşı	Türkisch	i ohne Punkt
İ	İrfan	Türkisch	I mit Punkt

Tabelle 23: Suffix-Heuristiken für die semiautomatische POS-Vorauszeichnung spanischer Lemmata (Auswahl)

Suffix	**Auszeichnung**	**Beispiel**
-al	N:ms / A:ms	*estatal*
-ano	N:ms / A:ms	*escribano*, *lejano*
-ario	N:ms / A:ms	*boticario*, *imaginario*
-astro	N:ms	*poetastro*
-azo	N:ms / A:ms	*perrazo*, *tontazo*
-cita	N:fs	*mujercita*
-cito	N:ms	*solecito*
-dad	N:fs	*verdad*
-dor	N:ms / A:ms	*colador*, *bebedor*
-ería	N:fs	*panadería*
-ero	N:ms / A:ms	*cajero*
-ía	N:fs	*energía*
-idad	N:fs	*identidad*
-ismo	N:ms	*socialismo*
-ista	N:ms / A:ms	*florista*
-iscar	V	*mordiscar*
-itis	N:fs	*bronquitis*
-logo	N:ms	*criminólogo*
-logos	N:mp	*filólogos*
-nte	N:ms / A:ms	*ayudante*, *refrescante*
-oide	N:ms	*sentimentaloide*
-ón	N:ms	*botellón*
-ona	N:fs	*mujerona*
-orio	N:ms / A:ms	*velorio*, *absolutorio*
-osidad	N:fs	*generosidad*
-ote	N:ms	*librote*
-tad	N:fs	*voluntad*

Tabelle 24: Präfix-Heuristiken für die semiautomatische POS-Vorauszeichnung spanischer Lemmata (Auswahl)

Präfix	**Beispiel**
ante-	*antepasado*
anti-	*antibiótico*
archi-	*archienemigo*
astro-	*astrofísica*
auto-	*automóvil*
bi-	*bilingüe*
co-	*cooperar*
contra-	*contradecir*
equi-	*equivalente*
ex-	*exnovio*
geo-	*geografía*
giga-	*gigabyte*
hiper-	*hipertensión*
hipo-	*hipotermia*
infra-	*infraestructura*
meta-	*metabolismo*
multi-	*multinacional*
pen-	*penúltimo*
peri-	*perímetro*
pluri-	*plurilingüe*
pos-	*posmoderno*
pro-	*promover*
proto-	*prototipo*
pseudo-	*pseudociencia*
re-	*reconstruir*
retro-	*retroceder*
super-	*supermercado*
uni-	*unilateral*

Tabelle 25: Morphologische False Friends bei Suffix-Heuristiken (Auswahl)

Suffix	**Typische Fehlklassifikation**	**Beispiel(e)**	**Hinweis/Gegenmaßnahme**
-a	N:fs statt A/V-Form	*alta*, *corta*	Sehr unspezifisch, nicht als N heuristisch taggen.
-ita	N:fs Dim. statt A/Partizip	*favorita*, *prescrita*, *escrita*	Dim.-Heuristik nur bei klarer Basis, Partizipien beachten.
-cita	N:fs Dim. statt Verbsequenz	*necesita*	Verbformen ausschließen, echte Diminutive Stamm+*-cita*.
-al	N statt A	*industrial* (A), *hospital* (N)	Ambig., N/A doppelt oder N nur bei Plural -ales.
-ano	N Gentilicio vs A	*mexicano*	N/A doppelt, sonst A priorisieren.
-ario	N vs A	*horario* (N), *secundario* (A)	Beide zulassen, Distribution prüfen.
-dor	N Agens vs A partizipial	*trabajador* (N/A), *refrescador* (A)	N/A doppelt, Geräte oft N.
-ero	N Beruf/Behälter vs A	*cajero* (N)	N/A doppelt, produktive N-Muster bevorzugen.
-ía	N:fs vs Verb Imperfekt	*energía*, *comía*	Akzent beachten, Verbformen ausschließen.
-ista	N Person vs A	*pianista*	N/A doppelt, Genus im Sg. invariant.
-nte	N Agens vs A	*estudiante* (N), *interesante* (A)	Ambig., N/A doppelt.
-ón	N Augmentativ vs Lexem	*problemón*, *camión*	Basis prüfen, nicht pauschal.
-ona	N Augmentativ vs Lexem	*mujerona*, *tizona*	Wie -ón.
-ote	N Augmentativ vs Lexem	*librote*, *bigote*	Augmentativ nur bei transparenter Basis.
-itis	N Medizin vs umgangssprachlich	*bronquitis*, *compritis*	Beide N, Stil markieren optional.
-oide	N vs A	*humanoide*	Ambig., N/A doppelt.
-orio	N vs A	*velorio*, *absolutorio*	Distribution nutzen.
-logo	N Beruf vs *logo*	*neurólogo*, *logo*	Akzent unterscheidet.

Fortsetzung auf der nächsten Seite

Suffix	Typische Fehlklassifikation	Beispiel(e)	Hinweis/Gegenmaßnahme
-osidad	N Abstraktum	*generosidad*	Sicheres N.
-dad/-tad	N Abstraktum	*verdad*, *voluntad*	Sicheres N.

Tabelle 26: Morphologische False Friends bei Präfix-Heuristiken (Auswahl)

Präfix	**Typische Fehlklassifikation**	**Beispiel(e)**	**Hinweis/Gegenmaßnahme**
super-	N Kompositum vs Adv/A Intensiv	*supermercado*, *superbien*	Als POS-Heuristik unscharf, Basis-POS prüfen.
re-	Verbpräfix vs adjektivisch intensiv	*reconstruir*, *rebueno*	Varietätenfilter, Basislemma prüfen.
ex-	N Ex-X vs Bindestrichartefakte	*exministro*, *ex ministro*	Hyphenation normalisieren, NE-Umfeld beachten.
co-	Verb/Nomen vs Ko-Autor mit Bindestrich	*cooperar*, *coautor/co-autor*	Bindestriche als Gremlins behandeln.
proto-	N vs A	*prototipo*, *protohistórico*	Ableitungstyp beachten.
pseudo-	N/A vs Orthographievarianten	*pseudociencia*, *pseudo-científico*	Varianten normalisieren.
pos-/post-	Orthographievariante	*posmoderno*, *postmoderno*	Normalform festlegen und erkennen.
multi-	N vs A	*multinacional*	Ambig., N/A doppelt.
archi-	A Intensiv vs N	*archifamoso*, *archienemigo*	Basis entscheiden lassen.
infra-	N vs V	*infraestructura*, *infravalorar*	Ohne Basis keine POS-Heuristik.

B Abdeckung der POS-Muster

Für den Hauptlauf auf dem EuroParl-v7-Korpus (vgl. Abschnitt 10.2) wurden mit den LEMoN-Grammatiken alle modellierten POS-Muster identifiziert. Tabelle 27 listet die Muster nach ihrer Häufigkeit. Hinweis: Der Graph zur Erkennung von `AN` war im Hauptlauf bewusst deaktiviert. Eine Gegenprobe mit aktiviertem `AN`-Graph ist in Ablationslauf L2 dokumentiert (vgl. Abschnitt 10.4.2). Die Beispielbelege sind so im Korpus gefunden und in `LEMoN.dic` kodifiziert. Groß-/Kleinschreibung und Akzentuierung wurden nicht normalisiert.

Kurzprofil der Verteilung: Alle 55 modellierten Muster sind belegt. Die beiden 2-MI-Muster `NN` und `NA` decken zusammen 44,36 % aller Matches ab. Die Top fünf Muster decken 64,77 %, die Top zehn 78,22 %, die Top zwanzig 90,36 %. Nach Längen ergibt sich: 2 MI 44,36 %. 3 MI 29,89 %. 4 MI 15,58 %. 5 MI 10,17 %.

Tabelle 27: Nominale POS-Muster (Längen 2–5) mit Frequenzen.

Nr	Main Items	POS	Freq	Prozent	Prozent kumulativ	Beispiel
1	2	NN	1 850 406	26,08 %	26,08 %	red de antenas
2	2	NA	1 297 115	18,28 %	44,36 %	Conflictos Armados
3	3	NNN	629 250	8,87 %	53,23 %	sistemas de acumulación de pensiones
4	3	NNA	448 201	6,32 %	59,55 %	concentraciones de ozono admisibles
5	3	NAN	370 479	5,22 %	64,77 %	teoría clásica de la burocracia
6	3	ANN	248 112	3,50 %	68,27 %	innecesaria adición de vitaminas
7	4	NNNN	216 892	3,06 %	71,33 %	Abogados Generales del Tribunal de Justicia
8	3	NAA	182 730	2,58 %	73,91 %	sector telefónico americano
9	3	ANA	155 765	2,20 %	76,11 %	gran jefe religioso
10	4	NNNA	150 654	2,12 %	78,22 %	Directiva sobre fuentes de energía renovables
11	4	NANN	142 922	2,01 %	80,23 %	sindicato mundial de editores de música
12	4	NNAN	117 393	1,65 %	81,88 %	granja de cerdos situada en el terreno
13	4	NANA	96 928	1,37 %	83,25 %	países europeos de la cuenca mediterránea
14	5	NNNNN	92 736	1,31 %	84,56 %	Asociación de Defensa de los Usuarios de Sangre de Portugal
15	4	ANNN	87 776	1,24 %	85,80 %	baja calidad del semen del hombre
16	3	AAN	86 361	1,22 %	87,02 %	nuevo Alto Representante
17	5	NANNN	60 683	0,86 %	87,88 %	programa especial para la prevención del Sida
18	5	NNNNA	60 654	0,85 %	88,74 %	mecanismos de intercambio de información sobre los antecedentes penales
19	4	NNAA	59 576	0,84 %	89,58 %	Comité de Seguridad Alimentaria Mundial
20	5	NNANN	56 825	0,80 %	90,38 %	inmunidad de los funcionarios encargados de la creación de una organización
21	4	ANNA	53 438	0,75 %	91,13 %	elevada incidencia de enfermedades gastrointestinales
22	5	NNNAN	49 920	0,70 %	91,83 %	Asamblea sobre la reforma de la Política Común de Pesca

Nr	Main Items	POS	Freq	Prozent	Prozent kumulativ	Beispiel
23	4	NAAN	43 882	0,62 %	92,45 %	inspecciones anuales obligatorias de las embarcaciones
24	4	ANAN	43 482	0,61 %	93,06 %	elevada calidad de la asistencia sanitaria
25	5	NANNA	39 934	0,56 %	93,62 %	acuerdo global de cooperación sobre el sistema económico
26	5	ANNNN	37 508	0,53 %	94,15 %	nuevo programa de acción en el campo de la juventud
27	5	NNANA	35 532	0,50 %	94,65 %	marca de confianza europea para el comercio electrónico
28	5	NANAN	32 639	0,46 %	95,11 %	Carta Comunitaria de los Derechos Fundamentales de los Trabajadores
29	4	AANN	32 185	0,45 %	95,56 %	posible nueva ocupación del edificio
30	5	ANNNA	24 594	0,35 %	95,91 %	compacto sistema de referencia de la comunidad internacional
31	5	NNNAA	23 110	0,33 %	96,24 %	mecanismos de notificación para la dotación indicativa asignada
32	5	NAANN	21 520	0,30 %	96,54 %	plan europeo común para la política del agua
33	5	ANANN	21 211	0,30 %	96,84 %	mayor banco público de crédito del mundo
34	5	ANNAN	20 509	0,29 %	97,13 %	mayor calidad de comercio electrónico en Europa
35	4	NAAA	20 250	0,29 %	97,42 %	política exterior común europea
36	5	NNAAN	19 447	0,27 %	97,69 %	antibióticos de uso médico derivados de la penicilina
37	4	AANA	18 835	0,27 %	97,96 %	única segunda lengua europea
38	5	NANAA	15 528	0,22 %	98,18 %	comisión permanente de la Asamblea Parlamentaria Paritaria
39	5	AANNN	13 950	0,20 %	98,38 %	único gran espacio de comercio sin barreras
40	5	NAANA	13 634	0,19 %	98,57 %	compañía energética monopolística de propiedad estatal
41	5	ANANA	13 093	0,18 %	98,75 %	nuevos Estados federados de la Alemania oriental
42	4	ANAA	11 590	0,16 %	98,91 %	nuevos Estados federados alemanes
43	5	NNAAA	9 672	0,14 %	99,05 %	régimen de garantía financiera obligatorio e inmediato
44	5	ANNAA	9 488	0,13 %	99,18 %	antiguo canciller de la República Federal Alemana
45	4	AAAN	9 311	0,13 %	99,31 %	-
46	5	AANNA	9 240	0,13 %	99,44 %	nueva Alta Comisaria para los Derechos Humanas

Nr	Main Items	POS	Freq	Prozent	Prozent kumulativ	Beispiel
47	5	AANAN	7 679	0,11 %	99,55 %	falsas e hipócritas declaraciones humanitarias de apoyo
48	5	ANAAN	7 278	0,10 %	99,66 %	complejo contexto electoral reinante en Europa
49	5	NAAAN	6 780	0,10 %	99,76 %	Agenda internacional específica para la próxima década
50	5	AAANN	4 757	0,07 %	99,83 %	próxima e inminente nueva propuesta de modificación
51	5	AANAA	3 661	0,05 %	99,88 %	limitada e incierta liberalización energética concedida
52	5	ANAAA	3 046	0,04 %	99,92 %	oscura maniobra política interior italiana
53	5	AAANA	2 971	0,04 %	99,95 %	-
54	5	NAAAA	2 687	0,04 %	99,99 %	Autoridad europea alimentaria eficaz e independiente
55	5	AAAAN	1 454	0,02 %	100,00 %	-

C Terminologische Vielfalt mit Belegen

Die folgenden Listen sind im Zuge der Literaturrecherche entstanden und fortlaufend weiterentwickelt worden. Sie versammeln Begriffe aus verschiedenen Fachrichtungen wie theoretische und allgemeine Linguistik, Lexikologie, Lexikographie, Psychologie, Psycholinguistik, Phraseologie, Kommunikationsforschung, Neurologie, Computer- und Korpuslinguistik, Informatik, Literaturwissenschaft und Sprachlehrforschung. Erfasst sind Bezeichnungen in Deutsch, Englisch und Spanisch, die auf MWE im Allgemeinen oder auf Teilklassen verweisen. Die Verwendung der Termini ist kontextabhängig und hängt von Disziplin, Theorie und Autorenschaft ab. Aus den Bezeichnungen lassen sich daher keine verbindlichen Taxonomien ableiten. Manche Begriffe sind disziplinspezifisch, andere an bestimmte Schulen gebunden. Die Nebeneinanderstellung impliziert weder Synonymie noch Gleichrangigkeit. Die Listen dienen vor allem als Arbeits- und Suchhilfe. Sie erheben keinen Anspruch auf Vollständigkeit und verdeutlichen zugleich, wie anspruchsvoll eine erschöpfende Bibliographie zu MWE ist. Die Quellen sind pragmatisch gewählte Belege und keine systematische Referenz für Erst- oder Hauptverwendungen. Viele sind sekundär, entstammen selbst Übersichtsarbeiten und bestehenden Sammlungen terminologischer Vielfalt. Schreib- und Bindestrichvarianten wurden in der Originalsprache belassen. Mögliche Bedeutungsüberlappungen werden hier nicht aufgelöst. Die Auswahl mischt fachsprachliche Termini mit kolloquialen Bezeichnungen.

C.1 Deutsche Terminologie

Autonomes Syntagma [E. Donalies 1994]

Automatisierter Redeteil [E. Donalies 1994]

Cluster [E. Donalies 1994]

Erstarrte Fügung [E. Donalies 1994]

Fertigbauteil [E. Donalies 1994]

Fertig vorhandene geprägte Wortverbindung [E. Donalies 1994]

Feste Fügung [C. Kunze 2007]

Feste Verbindung [E. Donalies 1994]

Feste Wendung [E. Donalies 1994]

Festgeprägter Satz [E. Donalies 1994]

Festes Syntagma [E. Donalies 1994]

Feststehende Redewendung [E. Donalies 1994]

Fixiertes Wortgefüge [E. Donalies 1994]

Floskel [E. Donalies 1994]

Formel [E. Donalies 1994]

Frasmus [E. Donalies 1994]

Gänzlich erstarrte Wortfügung [E. Donalies 1994]

Gebrauchsmetapher [E. Donalies 1994]

Geformter Wortblock [E. Donalies 1994]

Gemeinplatz [E. Donalies 1994]

Idiom [E. Donalies 1994]

Idiomatisch kombinierte Ausdrücke [C. Kunze 2007]

Idiomatische Lexemkette [E. Donalies 1994]

Idiomatische Phrase [E. Donalies 1994]

Idiomatische Redewendung [E. Donalies 1994]

Idiomatischer Phraseologismus [E. Donalies 1994]

Idiomatische Wendung [B. Pöll 2002]

Idiotismus [E. Donalies 1994]

Klischee [E. Donalies 1994]

Kodierte mehrmonematische Einheit [F. J. Hausmann 2007]

Kollokation [B.-B. Bischof 2007]

Komplexe Einheit [E. Donalies 1994]

Komplexe lexikalische Einheit [L. Lemnitzer 1997]

Komplexe Lexikoneinheit [L. Lemnitzer 1997]

Komplexes Lexem [L. Lemnitzer 1997]

Komplexe terminologische Einheit [L. Lemnitzer 1997]

Kookkurrenz [E. Donalies 1994]

Leerformel [E. Donalies 1994]

Lexikalische Solidarität [E. Coseriu 1967]

Mehrgliedrige lexikalische Einheiten [C. Kunze 2007]

Mehrgliedrige Komposita [C. Kunze 2007]

Mehrwort [G. Gréciano 2001]

Mehrwortausdruck [K.-U. Carstensen 2009]

Mehrworteinheit [L. Lemnitzer 1997]

Mehrwortgefüge [A. Stojić 2012]

Mehrwortlexem [C. Kunze 2007]

Mehrwortige Lexeme [C. Kunze 2007]

Mehrwortphänomen [E. Donalies 1994]

Mehrwortterm [D. Kim 2008]

Mehrworttermini [A. Caro Cedillo 2004]

Mehrwortverbindung [A. Stojić 2012]

Paralexem [A. Stojić 2012]

Phrase [E. Donalies 1994]

Phrasem [E. Donalies 1994]

Phraseolexem [E. Donalies 1994]

Phraseologische Einheit [E. Donalies 1994]

Phraseologismus [E. Donalies 1994]

Phraseoterm [F. J. Hausmann 2007]

Polylexem [A. Stojić 2012]

Redensart [E. Donalies 1994]

Redeweise [E. Donalies 1994]

Redewendung [E. Donalies 1994]

Routineformel [I. Sosa Mayor 2006]

Satzlexem [E. Donalies 1994]

Schlagwort [E. Donalies 1994]

Semantische Kongruenz [E. Donalies 1994]

Sinnkoppelung [E. Donalies 1994]

Sprachklischee [E. Donalies 1994]

Sprachlicher Schematismus [E. Donalies 1994]

Sprichwörtlich [E. Donalies 1994]

Stehende Redewendung [E. Donalies 1994]

Stereotyp [E. Donalies 1994]

Syntagma [E. Donalies 1994]

Syntaktische Gruppe [E. Donalies 1994]

Topos [E. Donalies 1994]

Usuelle Einheit des Sprachschatzes [E. Donalies 1994]

Verbale Stereotype [A. Stojić 2012]

Verknüpfungsrelation [E. Donalies 1994]

Vorgefertigte Einheit [E. Donalies 1994]

Vorgeformte Sprachwendung [E. Donalies 1994]

Wendung [E. Donalies 1994]

Wesenhafte Bedeutungsbeziehung [W. Porzig 1934]

Wortgefüge [E. Donalies 1994]

Wortgruppenlexem [E. Donalies 1994]

Wortverbindung [E. Donalies 1994]

Wortverknüpfung [E. Donalies 1994]

Wortzusammenstellung [E. Donalies 1994]

C.2 Englische Terminologie

Cliché [B. Kirkpatrick 1999]

Colligational pattern [D. Siepmann 2006]

Collocation [I. A. Bolshakov 2004]

Complex lexical unit [J. L. Cifuentes Honrubia 2011]

Complex terms [A. Doucet 2008]

Compound [R. Girju 2005]

Compound functional expression [T. Utsuro 2007]

Compound term [M. Gross 1986]

Compound utterance [M. Gross 1986]

Compound word [C. Krstev 2006]

Fixed expression [M. B. Villada Moirón 2005]

Frozen expression [D. Català 2007]

Frozen phrase [R. Kese 1992]

Idiom [J. Ayto 2010]

Idiomatic combination [A. Fazly 2006]

Idiomatic construction [J. Mateu 2007]

Idiomatic expression [P. Cook 2007]

Idiomatic patterning [M. Almela Sánchez 2006]

Idiomatic phrases [B. R. Haskell 2006]

Institutionalized phrases [I. A. Sag 2002]

Known units [V. Seretan 2011]

Lexical atom [F. Schmidt 2001]

Lexicalized phrases [I. A. Sag 2002]

Linguistic unit [F. Schmidt 2001]

Multi-element compound [K. T. Frantzi 1996]

Multi-lexemic expression [F. Guenthner 2004]

Multimember lexical units [J. A. Leoni de León 2008]

Multi-word [V. Arranz 2005]

Multi-word collocation [V. Seretan 2004]

Multi-word construction [J. Ayto 2010]

Multi-word expression [M. A. Attia 2006]

Multi-word lexeme [F. Segond 1996]

Multi-word lexical item [T. Baldwin 2010]

Multi-word lexical unit [G. Dias 1999]

Multi-word phrase [A. Nandi 2007]

Multi-word term [F. Schmidt 2001]

Multi-word terminological unit [F. Schmidt 2001]

Multi-word unit [A. Savary 2005]

N-gram [F. Schmidt 2001]

Phrase [A. Nandi 2007]

Phraseological unit [H. Alisoy 2025]

Phraseologism [D. Anastasiou 2010]

Polylexical unit [D. Maurel 2005]

Polylog [V. Seretan 2011]

Polywords [J. D. Becker 1975]

Set phrase [I. Mel'čuk 1998]

Sticky phrase [F. Schmidt 2001]

Structurally complex unit [D. Siepmann 2006]

Term [F. Schmidt 2001]

Thematic fusion [J. A. Leoni de León 2008]

Verb–noun collocation [S. Venkatapathy 2005]

Word-group lexemes [E. Luis Álvarez 2019]

C.3 Spanische Terminologie

Adagio [M. T. Zurdo 2007]

Colocación [A. Martín Bosque 2008]

Construcción multiverbal [J. Martínez Marín 2000]

Complejo verbal [J. Martínez Marín 2000]

Compuesto sintagmático [J. Martínez Marín 2000]

Decir(es) [M. T. Zurdo 2007]

Dicho [M. T. Zurdo 2007]

Enunciado fraseológico [M. T. Zurdo 2007]

Expresión [M. T. Zurdo 2007]

Expresión fija [G. Mapelli 2004]

Expresión fraseológica [M. T. Zurdo 2007]

Expresión idiomática [G. Mapelli 2004]

Expresión multilexémica [I. Mel'čuk 2001]

Expresión multipalabra [I. Nica 2004]

Expresión pluriverbal [M. T. Zurdo 2007]

Expresión poliléxica [M. T. Zurdo 2007]

Fórmula [M. T. Zurdo 2007]

Fórmula de conversación [Y. I. Hernández Muñoz 2019]

Fórmula pluriverbal [M. T. Zurdo 2007]

Frase hecha [G. Mapelli 2004]

Frase proverbial [M. T. Zurdo 2007]

Frase sentenciosa [M. T. Zurdo 2007]

Frasema [I. Mel'čuk 2001]

Fraseolexema [M. T. Zurdo 2007]

Fraselogismo [M. T. Zurdo 2007]

Giro [G. Mapelli 2004]

Idiom [G. Mapelli 2004]

Lexía [G. Mapelli 2004]

Locución [G. Mapelli 2004]

Locución coordinada [A. Ricós Vidal 2008]

Modismo [G. Mapelli 2004]

Modo de decir [B. Pöll 2002]

Multipalabra [J. Herrera 2005]

Palabra cliché [M. Alonso Ramos 2004]

Palabra compleja [J. Martínez Marín 2000]

Palabra compuesta [L. Hierrezuelo Sabatela 2006]

Paremia [C. Jarillot Rodal 2010]

Pluriverbal [E. T. Montoro del Arco 2021]

Proverbio [M. T. Zurdo 2007]

Perífrase léxica [G. Mapelli 2004]

Refrán [M. T. Zurdo 2007]

Sintagma complejo [A. Ricós Vidal 2008]

Sintagma estereotipado [G. Mapelli 2004]

Término multipalabra [L. A. Barrón 2006]

Unidad de texto repetido [M. T. Zurdo 2007]

Unidad fraseológica [G. Mapelli 2004]

Unidad léxica compleja [M. Alonso Ramos 2004]

Unidad léxica compuesta [C. Timm 2010]

Unidad léxica multipalabra [M. Lloberes 2013]

Unidad léxica polilexemática [H. Provencio Garrigós 2003]

Unidad léxica poliparadigmática [M. L. Izquierdo Guzmán 1992]

Unidad multilexémica [Y. I. Hernández Muñoz 2019]

Unidad multipalabra [M. Lloberes 2013]

Unidad pluriverbal [J. Martínez Marín 2000]

Unidad sintagmática [J. Martínez Marín 2000]

Unidad terminológica multilexémica [M. R. Vilhena Costa 2001]

Unidad terminológica poliléxica [R. Estopà Bagot 2001]

Unidad terminológica polilexemática [C. Mellado-Blanco 2008]

Literatur

[S. Abalada 2009] Silvana Abalada, Vera Cabarrão & Aida Cardoso. *Proposta de Classificação Semânticade Unidades Lexicais Multipalavra Nominais.* In: Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística, Seiten 81–94. APL, 2009.

[J. Albrecht 2000] Jörn Albrecht. Europäischer Strukturalismus: ein forschungsgeschichtlicher Überblick. Francke, Tübingen / Basel, 2. Auflage, 2000.

[S. Alghazo 2022] Sharif Alghazo, Imran Alrashdan, Mohd Nour Al Salem & Essa Salem. *Idioms as Pragmatic Messages.* Dirasat: Human and Social Sciences, Bd. 49, Nr. 4, Seiten 425–442, Juli 2022.

[H. Alisoy 2025] Hasan Alisoy. *A Structural and Semantic Classification of Phraseological Units in English.* Global Spectrum of Research and Humanities, Bd. 2, Nr. 3, 2025.

[M. Almela Sánchez 2006] Moisés Almela Sánchez (Hg). From words to lexical units, Bd. 35 von *Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation.* Peter Lang, Frankfurt am Main, 2006.

[M. Alonso Ramos 2004] Margarita Alonso Ramos. Las construcciones con verbo de apoyo. Visor Libros, Madrid, 2004.

[M. Alonso Ramos 2010] Margarita Alonso Ramos, Alfonso Nishikawa & Orsolya Vincze. *DiCE in the web: An online Spanish collocation dictionary.* In: Sylviane Granger & Magali Paquot (Hgg), eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Bd. 7 von *Cahiers du CENTAL*, Seiten 369–374, Louvain-la-Neuve, 2010. Presses universitaires de Louvain.

[D. Anastasiou 2010] Dimitra Anastasiou. Idiom treatment experiments in machine translation. Cambridge Scholars Publishing, Newcastle upon Tyne, 2010.

[V. Arranz 2005] Victoria Arranz, Jordi Atserias & Mauro Castillo. *Multiwords and Word Sense Disambiguation.* In: CICLing'05: Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing, Seiten 250 – 262. Springer-Verlag Berlin, Heidelberg, 2005.

[M. A. Attia 2006] Mohammed A. Attia. *Accommodating Multiword Expressions in an Arabic LFG Grammar.* In: Advances in Natural Language Processing (FinTAL 2006), Bd. 4139 von *Lecture Notes in Computer Science*, Seiten 87–98, Finland, 2006. Springer.

[J. Ayto 2010] John Ayto. Idioms, Seiten 404–407. Elsevier, Amsterdam, 2010.

[S. Bagdasarov 2024] Sergei Bagdasarov & Elke Teich. *Multi-word expressions in biomedical abstracts and their plain English adaptations.* In: Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH), Seiten 483–488, 2024.

[T. Baldwin 2010] Timothy Baldwin & Su Nam Kim. *Multiword Expressions.* In: Nitin Indurkhya & Fred J. Damerau (Hgg), Handbook of Natural Language Processing, Seiten 267–292. Chapman and Hall/CRC, Boca Raton, 2. Auflage, 2010.

[J. Baptista 2003] Jorge Baptista. *Some Families of Compound Temporal Adverbs in Portuguese.* In: Proceedings of the Workshop on Finite-State Methods for Natural Language Processing, 2003.

[L. A. Barrón 2006] L. Alberto Barrón, Gerardo Sierra & Elio Villaseñor. *C-value aplicado a la extración de términos multipalabra en documentos técnicos y científicos en español.* In: 7th Mexican International Conference on Computer Science (ENC 2006). IEEE Computer Press, 2006.

[A. Barrón-Cedeño 2009] Alberto Barrón-Cedeño, Gerardo Sierra, Patrick Drouin & Sophia Ananiadou. *An Improved Automatic Term*

Recognition Method for Spanish. In: Alexander Gelbukh (Hg), Computational Linguistics and Intelligent Text Processing: 10th International Conference, CICLing 2009, Mexico City, Mexico. Proceedings, Seiten 125–136, Berlin, Heidelberg, 2009. Springer.

[L. Bauer 1983] Laurie Bauer. English Word-Formation. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 1983.

[J. D. Becker 1975] Joseph D. Becker. *The Phrasal Lexicon.* In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing (TINLAP '75), Seiten 60–63, Cambridge, Massachusetts, USA, 1975. Association for Computing Machinery.

[C. Beiras Cunqueiro 2005] Cibrán Beiras Cunqueiro. Lösung von lexikalischen Ambiguitäten in der spanischen Sprache mittels des Formalismus Elag. Abschlussarbeit, Ludwig-Maximilians-Universität München, München, 2005.

[E. J. L. Bell 2007] Edward J. L. Bell. *Collocation Statistical Analysis Tool: An evaluation of the effectiveness of extracting domain phrases via collocation.* Dissertation, Lancaster University, 2007.

[S. Bird 2009] Steven Bird, Ewan Klein & Edward Loper. Natural Language Processing with Python. O'Reilly Media, Sebastopol, 1. Auflage, 2009.

[B.-B. Bischof 2007] Beatrice-Barbara Bischof. *Französische Kollokationen diachron. Eine korpusbasierte Analyse.* Dissertation, Universität Stuttgart, Institut für Linguistik/Romanistik, 2007.

[X. Blanco Escoda 2001] Xavier Blanco Escoda. *Les dictionnaires électroniques de l'espagnol (DELASs et DELACs).* Lingvisticae Investigationes, Bd. 23, Nr. 2, Seiten 201–218, 2001. In der Publikation als: Xavier Blanco.

[X. Blanco Escoda 2002] Xavier Blanco Escoda. *Les déterminants figés.* Langages, Bd. 36, Nr. 145, Seiten 61–81, 2002. In der Publikation als: Xavier Blanco.

[X. Blanco Escoda 2008] Xavier Blanco Escoda. *Variantes diatópicas de las locuciones nominales endocéntricas en español.* In: Actas del XV Congreso Internacional de la Asociación de Lingüística y Filología de América Latina (ALFAL), Montevideo, 2008.

[X. Blanco Escoda 2010] Xavier Blanco Escoda. *Etiquetas semánticas de hecho como género próximo en la definición lexicográfica.* Quaderns de Filologia. Estudis lingüístics, Nr. 15, Seiten 159–178, 2010.

[L. Bloomfield 1933] Leonard Bloomfield. Language. Henry Holt and Company, New York, 1933.

[I. A. Bolshakov 2004] Igor A. Bolshakov & Sabino Miranda-Jiménez. *A Small System Storing Spanish Collocations.* In: Alexander F. Gelbukh (Hg), Computational Linguistics and Intelligent Text Processing, Bd. 2945 von *Lecture Notes in Computer Science*, Seiten 248–252. Springer, Berlin, Heidelberg, 2004.

[P. F. Brown 1990] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin. *A Statistical Approach to Machine Translation.* Computational Linguistics, Bd. 16, Nr. 2, Seiten 79–85, 1990.

[C. Buenafuentes de la Mata 2007] Cristina Buenafuentes de la Mata. *Procesos de gramaticalización y lexicalización en la formación de compuestos en Español.* Dissertation, Universidad Autónoma de Barcelona, Bellaterra, 2007.

[C. Buenafuentes de la Mata 2017] Cristina Buenafuentes de la Mata. *Sobre la delimitación entre compuestos sintagmáticos y locuciones: Nuevas aportaciones desde la diacronía.* Hispania, Bd. 100, Nr. 4, Seiten 568–579, 2017.

[C. Buenafuentes de la Mata 2021] Cristina Buenafuentes de la Mata. *Main compounding types in Spanish: synchronic issues.* In: Antonio Fábregas, Víctor Acedo-Matellán, Grant Armstrong, María Cristina Cuervo & Isabel Pujol Payet (Hgg), The Routledge Handbook of Spanish Morphology, Routledge Spanish Language Handbooks, Seiten 285–302. Routledge, London/New York, 2021.

[H. Burger 2010] Harald Burger. Phraseologie. Eine Einführung am Beispiel des Deutschen, Bd. 36 von *Grundlagen der Germanistik*. Erich Schmidt Verlag, Berlin, 4. Auflage, 2010.

[H. Burger 1982] Harald Burger, Annelies Buhofer & Ambros Sialm (Hgg). Handbuch der Phraseologie. Walter de Gruyter, Berlin; New York, 1982.

[N. Calzolari 1991] Nicoletta Calzolari. *Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base.* In: Uri Zernik (Hg), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Seiten 191–208. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[A. Caro Cedillo 2004] Ana Caro Cedillo. Fachsprachliche Kollokationen. Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch, Bd. 63 von *Forum für Fachsprachen-Forschung*. Gunter Narr Verlag, Tübingen, 2004.

[K.-U. Carstensen 2009] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne J. Jekat, Ralf Klabunde & Hagen Langer (Hgg). Computerlinguistik und Sprachtechnologie. Eine Einführung. Spektrum Akademischer Verlag, Heidelberg, 3. Auflage, 2009.

[L. Castro 2025] Laura Castro & Marcos Garcia. *Gathering Compositionality Ratings of Ambiguous Noun-Adjective Multiword Expressions in Galician.* In: Atul Kr. Ojha, Voula Giouli, Verginica Barbu Mititelu, Mathieu Constant, Gražina Korvel, A. Seza Doğruöz & Alexandre

Rademaker (Hgg), Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), Seiten 32–40, Albuquerque, New Mexico, U.S.A., Mai 2025. Association for Computational Linguistics.

[D. Català 2007] Dolors Català & Jorge Baptista. *Spanish Adverbial Frozen Expressions.* In: Nicole Grégoire, Stefan Evert & Su Nam Kim (Hgg), Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Seiten 33–40. Association for Computational Linguistics, Prague, Czech Republic, Juni 2007.

[H.-R. Chae 2015] Hee-Rahk Chae. *Idioms: Formally Flexible but Semantically Non-transparent.* In: Hai Zhao (Hg), Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters, Seiten 46–54, Shanghai, China, Oktober 2015.

[N. Chomsky 1957] Noam Chomsky. Syntactic Structures, Bd. 4 von *Janua Linguarum. Series Minor.* Mouton, The Hague, 1957.

[N. Chomsky 1965] Noam Chomsky. Aspects of the Theory of Syntax. Nummer 11 in MIT Research Monographs. MIT Press, Cambridge, MA, 1965.

[J. L. Cifuentes Honrubia 2011] José Luis Cifuentes Honrubia & Susana Rodríguez Rosique (Hgg). Spanish Word Formation and Lexical Creation, Bd. 1 von *IVITRA Research in Linguistics and Literature.* John Benjamins Publishing Company, Amsterdam/Philadelphia, 2011.

[M. Constant 2003] Matthieu Constant & Anastasia Yannacopoulou. *Le dictionnaire électronique du grec moderne: conception et développement d'outils pour son enrichissement et sa validation.* In: Studies in Greek Linguistics, Proceedings of the 23rd annual meeting of the Department of Linguistics (2002), Bd. 2, Seiten 783–791, Thessaloniki, 2003. Faculty of Philosophy, Aristotle University of Thessaloniki.

[P. Cook 2007] Paul Cook, Afsaneh Fazly & Suzanne Stevenson. *Pulling their Weight: Exploiting Syntactic Forms for the*

Automatic Identification of Idiomatic Expressions in Context. In: Nicole Grégoire, Stefan Evert & Su Nam Kim (Hgg), Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Seiten 41–48, Prague, Czech Republic, Juni 2007. Association for Computational Linguistics.

[G. Corpas Pastor 1996] Gloria Corpas Pastor. Manual de fraseología española, Bd. 3 von *Biblioteca románica hispánica. Manuales.* Gredos, Madrid, 1996.

[E. Coseriu 1967] Eugenio Coseriu. *Lexikalische Solidaritäten.* Poetics, Bd. 1, Nr. 3, Seiten 293–312, 1967.

[COST 2012] COST. *Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1207: PARSEME: PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing.* Memorandum of Understanding approved by the COST Committee of Senior Officials at its 186th meeting, November 2012.

[B. Courtois 1990] Blandine Courtois & Max Silberztein. *Dictionnaires électroniques du français [liminaire].* Langue française, Bd. 87, Seiten 3–4, 1990.

[B. Daille 1994] Beatrice Daille. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology.* In: David D. McDonald (Hg), The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Seiten 29–36, Nonantum Inn, Kennebunkport, Maine, Juni 1994. Association for Computational Linguistics.

[B. Daille 2000] Beatrice Daille. *Morphological Rule Induction for Terminology Acquisition.* In: COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics, Seiten 215–221, Saarbrücken, Germany, August 2000.

[B. Daille 1996] Béatrice Daille, Benoît Habert, Christian Jacquemin & Jean Royauté. *Empirical observation of term variations and principles for their description.* Terminology, Bd. 3, Nr. 2, Seiten 197–257, 1996.

[A. M. Di Sciullo 1987] Anna Maria Di Sciullo & Edwin Williams. On the Definition of Word. Nummer 14 in Linguistic Inquiry Monographs. The MIT Press, Cambridge, MA/London, 1. Auflage, 1987.

[G. Dias 1999] Gaël Dias, Sylvie Guilloré & José Gabriel Pereira Lopes. *Multilingual Aspects of Multiword Lexical Units.* In: Špela Vintar (Hg), Proceedings of the Workshop Language Technologies — Multilingual Aspects, Seiten 11–21, Ljubljana, Slovenia, Juli 1999. Faculty of Arts, University of Ljubljana.

[D. Dobrovol'skij 2023] Dmitrij Dobrovol'skij. *Zum Wesen der Idiomatizität in der Phraseologie.* Verbum, Bd. 45, Nr. 1, Seiten 49–75, 2023.

[E. Donalies 1994] Elke Donalies. *Idiom, Phraseologismus oder Phrasem? Zum Oberbegriff eines Bereichs der Linguistik.* Zeitschrift für germanistische Linguistik, Bd. 22, Nr. 3, Seiten 334–349, 1994.

[A. Doucet 2008] Antoine Doucet & Miro Lehtonen. *Let's Phrase It: INEX Topics Need Keyphrases.* In: Proceedings of the ACM SIGIR 2008 Workshop on Focused Retrieval (Question Answering, Passage Retrieval, Element Retrieval), Seiten 9–14, Singapore, Juli 2008.

[A. Dufter 2017] Andreas Dufter & Elisabeth Stark (Hgg). Manual of Romance Morphosyntax and Syntax, Bd. 17 von *Manuals of Romance Linguistics.* De Gruyter Mouton, Berlin/Boston, 2017.

[A. Eisele 2010] Andreas Eisele & Yu Chen. *MultiUN: A Multilingual Corpus from United Nation Documents.* In: Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner &

Daniel Tapias (Hgg), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, Mai 2010. European Language Resources Association (ELRA).

[R. Estopà Bagot 2001] Rosa Estopà Bagot. *Extracción de terminología: elementos para la construcción de un extractor.* TradTerm, Bd. 7, Seiten 225–250, 2001.

[A. Fazly 2006] Afsaneh Fazly & Suzanne Stevenson. *Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations.* In: Diana McCarthy & Shuly Wintner (Hgg), 11th Conference of the European Chapter of the Association for Computational Linguistics, Seiten 337–344, Trento, Italy, April 2006. Association for Computational Linguistics.

[C. J. Fillmore 1988] Charles J. Fillmore, Paul Kay & Mary Catherine O'Connor. *Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone.* Language, Bd. 64, Nr. 3, Seiten 501–538, 1988.

[J. Y. Findlay 2019] Jamie Y. Findlay. *Multiword Expressions and the Lexicon.* Dissertation, University of Oxford, Oxford, Januar 2019.

[J. R. Firth 1957] J. R. Firth. Modes of Meaning, Seiten 190–215. Oxford University Press, London, 1957.

[K. T. Frantzi 1996] Katerina T. Frantzi & Sophia Ananiadou. *Extracting Nested Collocations.* In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, Seiten 41–46, Copenhagen, Denmark, August 1996. Association for Computational Linguistics.

[K. T. Frantzi 2000] Katerina T. Frantzi, Sophia Ananiadou & Hideki Mima. *Automatic recognition of multi-word terms: the C-value/NC-value method.* International Journal on Digital Libraries, Bd. 3, Nr. 2, Seiten 115–130, 2000.

[M. García-Page Sánchez 2008] Mario García-Page Sánchez. Introducción a la fraseología española. Estudio de las locuciones, Bd. 6 von *Autores,*

Textos y Temas. Lingüística. Anthropos Editorial, Barcelona, 2008.

[G. García Subies 2024] Guillem García Subies, Álvaro Barbero Jiménez & Paloma Martínez Fernández. *A comparative analysis of Spanish Clinical encoder-based models on NER and classification tasks.* Journal of the American Medical Informatics Association, Bd. 31, Nr. 9, Seiten 2137–2146, 2024.

[M. Geierhos 2006] Michaela Geierhos. Grammatik der Menschenbezeichner in biographischen Kontexten. Abschlussarbeit, Ludwig-Maximilians-Universität München, CIS, München, März 2006.

[R. Girju 2005] Roxana Girju, Dan Moldovan, Marta Tatu & Daniel Antohe. *On the semantics of noun compounds.* Computer Speech & Language, Bd. 19, Nr. 4, Seiten 479–496, 2005.

[Y. Graham 2019] Yvette Graham, Barry Haddow & Philipp Koehn. *Translationese in Machine Translation Evaluation*, Juni 2019.

[M. Grandury 2024] María Grandury. *The #Somos600M Project: Generating NLP resources that represent the diversity of the languages from LATAM, the Caribbean, and Spain*, Juli 2024.

[G. Gréciano 2001] Gertrud Gréciano. *Sprachfertigteile, ihre kognitive und kommunikative Performanz.* Internationale Kulturwissenschaften, 2001.

[J. Grimm 1854] Jacob Grimm & Wilhelm Grimm. *Vorrede.* In: Deutsches Wörterbuch, Bd. 1, Seiten Sp. I–XXVI. S. Hirzel, Leipzig, Mai 1854. Vorrede zum ersten Band.

[G. Gross 1996] Gaston Gross. Les expressions figées en français: noms composés et autres locutions. L'essentiel français. Ophrys, Paris/Gap, 1996.

[G. Gross 2012] Gaston Gross. Manuel d'analyse linguistique: Approche sémantico-syntaxique du lexique. Sens et structures.

Presses universitaires du Septentrion, Villeneuve d'Ascq, 2012.

[M. Gross 1979] Maurice Gross. *On the Failure of Generative Grammar.* Language, Bd. 55, Nr. 4, Seiten 859–885, 1979.

[M. Gross 1982] Maurice Gross. *Une classification des phrases « figées » du français.* Revue québécoise de linguistique, Bd. 11, Nr. 2, Seiten 151–185, 1982.

[M. Gross 1984] Maurice Gross. *Lexicon-Grammar and the Syntactic Analysis of French.* In: 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, Seiten 275–282, Stanford, California, USA, Juli 1984. Association for Computational Linguistics.

[M. Gross 1986] Maurice Gross. *Lexicon - Grammar The Representation of Compound Words.* In: Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics, Seiten 1–6, Bonn, Germany, August 1986.

[M. Gross 1988] Maurice Gross. *Les limites de la phrase figée.* Langages, Bd. 23, Nr. 90, Seiten 7–22, 1988.

[M. Gross 1989] Maurice Gross. *The Use of Finite Automata in the Lexical Representation of Natural Language.* In: Maurice Gross & Dominique Perrin (Hgg), Electronic Dictionaries and Automata in Computational Linguistics, Bd. 377 von *Lecture Notes in Computer Science*, Seiten 34–50, Berlin/Heidelberg, 1989. Springer.

[M. Gross 1993a] Maurice Gross. *Les phrases figées en français.* L'Information grammaticale, Nr. 59, Seiten 36–41, 1993.

[M. Gross 1993b] Maurice Gross. *Local grammars and their representation by finite automata.* In: Michael Hoey (Hg), Data, Description, Discourse: Papers on the English Language in Honour of John McH. Sinclair, Seiten 26–38. HarperCollins, London, 1993.

[M. Gross 1994] Maurice Gross. *Constructing Lexicon-Grammars.* In: B. T. S. Atkins & Antonio Zampolli (Hgg), Computational Approaches to the Lexicon, Seiten 213–264. Oxford University Press, Oxford, 1994.

[M. Gross 1995] Maurice Gross. *On Counting Meaningful Units in Texts.* In: Sergio Bolasco, Ludovic Lebart & André Salem (Hgg), Actes des 3es Journées internationales d'Analyse statistique des Données Textuelles (JADT 1995), Seiten 5–18, Rome, Italy, Dezember 1995. CISU.

[M. Gross 1999] Maurice Gross. *A bootstrap method for constructing local grammars.* In: Neda Bokan (Hg), Proceedings of the Symposium on Contemporary Mathematics, Seiten 229–250. University of Belgrade, Belgrade, 1999.

[F. Guenthner 1998] Franz Guenthner. *Constructions, classes et domaines : concepts de base pour un dictionnaire électronique de l'allemand.* Langages, Bd. 32, Nr. 131, Seiten 45–55, 1998.

[F. Guenthner 2005] Franz Guenthner. *Local Grammars in Corpus Calculus.* In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2005", Seiten 616–620, Moscow, Russia, Juni 2005.

[F. Guenthner 2004] Franz Guenthner & Xavier Blanco Escoda. *Multi-Lexemic Expressions: An overview.* In: Christian Leclère, Éric Laporte, Mireille Piot & Max Silberztein (Hgg), Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis & Lexicon-Grammar: Papers in honour of Maurice Gross, Bd. 24 von *Lingvisticae Investigationes Supplementa*, Seiten 239–252. John Benjamins, Amsterdam/Philadelphia, 2004.

[F. Guenthner 1994] Franz Guenthner & Petra Maier. *Das CISLEX-Wörterbuchsystem.* CIS-Berichte, Bd. 76, 1994.

[E. Gutiérrez Rubio 2024] Enrique Gutiérrez Rubio & Jana Plačková. *Introducción al estudio de la fraseología taurina recogida en* La

fraseología taurina... y su pícaro humor *de Manuel Bermejo Hernández*. ELUA: Estudios de Lingüística. Universidad de Alicante, Nr. 41, Seiten 185–200, 2024.

[J. Han 2018] Jaeho Han, Changhoe Hwang, Seongyong Choi, Gwanghoon Yoo, Éric Laporte & Jeesun Nam. *DECO-MWE: Building a Linguistic Resource of Korean Multiword Expressions for Feature-Based Sentiment Analysis*. In: Kiyoaki Shirai (Hg), Proceedings of the 13th Workshop on Asian Language Resources (ALR-13), Seiten 14–20, Miyazaki, Japan, Mai 2018. ELRA.

[R. A. Harris 1993] Randy Allen Harris. The Linguistics Wars. Oxford University Press, Oxford, 1993.

[Z. S. Harris 1963] Zellig S. Harris. Structural Linguistics. Nummer P52 in Phoenix Books. The University of Chicago Press, Chicago, Illinois, U.S.A., 1963.

[Z. S. Harris 1968] Zellig S. Harris. Mathematical Structures of Language. Nummer 21 in Interscience Tracts in Pure and Applied Mathematics. John Wiley & Sons, New York, 1968.

[Z. S. Harris 1991] Zellig S. Harris. A Theory of Language and Information: A Mathematical Approach. Clarendon Press, Oxford, 1991.

[B. R. Haskell 2006] Benjamin R. Haskell, Chandra Barnett & Christiane Fellbaum. *A Proposal for the Automatic Distinction of Homomorphic Idiomatic and Non-Idiomatic Phrases in WordNet*. In: Petr Sojka, Key-Sun Choi, Christiane Fellbaum & Piek Vossen (Hgg), Proceedings of the Third International WordNet Conference (GWC 2006), Seogwipo, Jeju Island, Korea, Januar 2006. KAIST.

[F. J. Hausmann 2007] Franz Josef Hausmann. *Die Kollokationen im Rahmen der Phraseologie - Systematische und historische Darstellung*. Zeitschrift für Anglistik und Amerikanistik, Bd. 55, Nr. 3, Seiten 217–234, 2007.

[W. He 2025] Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart & Aline Villavicencio. *Investigating*

Idiomaticity in Word Representations. Computational Linguistics, Bd. 51, Nr. 2, Seiten 505–555, Juni 2025.

[H. Heleine 2009] Henry Heleine. *Automatic Term Recognition.* BSc Internet Computing thesis, Department of Computer Science, The University of Manchester, Mai 2009. Supervisor: Sophia Ananiadou.

[Y. I. Hernández Muñoz 2019] Yaiza Irene Hernández Muñoz. *Las construcciones francesas fundamentales: definición y aplicación de una nueva unidad fraseológica.* Dissertation, Universidad Complutense de Madrid, Facultad de Filología, Madrid, 2019. Director: Álvaro Arroyo Ortega.

[J. Herrera 2005] Jesús Herrera, Anselmo Peñas & Felisa Verdejo. *Técnicas Aplicadas al Reconocimiento de Implicación Textual.* In: Roque Marín, Eva Onaindia, Alberto Bugarín & José Santos (Hgg), Actas del CAEPIA 2005. Asociación Española para la Inteligencia Artificial (AEPIA), 2005.

[L. Hierrezuelo Sabatela 2006] Lisset Hierrezuelo Sabatela, Sira Elena Palazuelos Cagigas, José Luis Martín Sánchez & Manuel R. Mazo Quintas. *Combinación de algoritmos de predicción de palabras y de expansiones en español para sistemas de ayuda a la escritura para personas con discapacidad.* In: Actas del IV Congreso Iberoamericano sobre Tecnologías de Apoyo a la Discapacidad (IBERDISCAP 2006), Seiten 29–34, Vitória, Brasil, Februar 2006.

[C. Hwang 2018] Changhoe Hwang, Gwanghoon Yoo & Jee-Sun Nam. *Constructing Korean Multi-Word Expression Dictionary DecoMWE for Feature-Based Sentiment Analysis.* Language & Information Society, Bd. 35, Seiten 407–458, 2018.

[ISO/TC 37/SC 4 2006] ISO/TC 37/SC 4. *Language resource management — Lexical markup framework (LMF).* Rapport technique ISO/TC 37/SC 4 N130 Rev.9, International Organization for Standardization (ISO), Technical Committee 37, Subcommittee 4, Geneva, 2006. Working document.

[M. L. Izquierdo Guzmán 1992] María Laura Izquierdo Guzmán. *Estudio léxico-semántico de los términos que delimitan tiempo en 'día'. Investigación diacrónica.* Tesis doctoral, Universidad de La Laguna, La Laguna, 1992. Director: Cristóbal Corrales Zumbado. Facultad de Filología, Departamento de Filología Española.

[C. Jarillot Rodal 2010] Cristina Jarillot Rodal (Hg). Bestandsaufnahme der Germanistik in Spanien: Kulturtransfer und methodologische Erneuerung. Peter Lang, Bern, 2010.

[S. Jiang 2020] Shengjun Jiang, Xin Jiang & Anna Siyanova-Chanturia. *The processing of multiword expressions in children and adults: An eye-tracking study of Chinese.* Applied Psycholinguistics, Bd. 41, Nr. 4, Seiten 901–931, 2020.

[J. S. Justeson 1995] John S. Justeson & Slava M. Katz. *Technical terminology: some linguistic properties and an algorithm for identification in text.* Natural Language Engineering, Bd. 1, Nr. 1, Seiten 9–27, März 1995.

[K. Kanclerz 2022] Kamil Kanclerz & Maciej Piasecki. *Deep Neural Representations for Multiword Expressions Detection.* In: Samuel Louvan, Andrea Madotto & Brielen Madureira (Hgg), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Seiten 444–453, Dublin, Ireland, Mai 2022. Association for Computational Linguistics.

[G. Katz 2006] Graham Katz & Eugenie Giesbrecht. *Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis.* In: Begoña Villada Moirón, Aline Villavicencio, Diana McCarthy, Stefan Evert & Suzanne Stevenson (Hgg), Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Seiten 12–19, Sydney, Australia, Juli 2006. Association for Computational Linguistics.

[R. Kese 1992] Ralf Kese, Friedrich Dudda & Marianne Kugler. *Automatic Proofreading of Frozen Phrases in German.*

In: COLING 1992 Volume 3: The 14th International Conference on Computational Linguistics, Seiten 1029–1033, Nantes, France, August 1992. Association for Computational Linguistics.

[D. Kim 2008] Daewoo Kim. *Semantische Analyse und automatische Gewinnung von branchenspezifischem Vokabular für E-Commerce.* Dissertation, Ludwig-Maximilians-Universität München, München, Januar 2008.

[B. Kirkpatrick 1999] Betty Kirkpatrick. Clichés: over 1500 phrases explored and explained. St. Martin's Griffin, New York, 1999.

[P. Koehn 2005] Philipp Koehn. *Europarl: A Parallel Corpus for Statistical Machine Translation.* In: Proceedings of Machine Translation Summit X: Papers, Seiten 79–86, Phuket, Thailand, September 2005.

[C. Krstev 2006] Cvetana Krstev & Duško Vitas. *Finite State Transducers for Recognition and Generation of Compound Words.* In: Tomaž Erjavec & Jerneja Žganec Gros (Hgg), Language Technologies, IS-LTC 2006: Proceedings of the 5th Slovenian and 1st International Conference, Seiten 192–197, Ljubljana, Slovenia, Oktober 2006. Institut Jožef Stefan.

[C. Kunze 2007] Claudia Kunze & Lothar Lemnitzer. Computerlexikographie. Eine Einführung. Narr Francke Attempto Verlag, Tübingen, 2007.

[A. Ladilova 2024] Anna Ladilova, Katharina Müller, Simone F. Gomes & Joachim Born. *Linguistic change in times of the COVID-19 pandemic: a corpus linguistic comparison of language contact phenomena in Romance languages.* Zeitschrift für romanische Philologie, Bd. 140, Nr. 1, Seiten 1–29, 2024.

[S. Langer 1997] Stefan Langer. *Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX.* Dissertation,

Ludwig-Maximilians-Universität München, CIS, München, 1997.

[S. Langer 1998] Stefan Langer. *Zur Morphologie und Semantik von Nominalkomposita.* In: Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 98), Seiten 83–97, Bonn, 1998.

[S. Langer 1996] Stefan Langer, Petra Maier & Jürgen Oesterle. *CISLEX: An Electronic Dictionary for German: Its Structure and a Lexicographic Application.* In: Proceedings of COMPLEX '96, Budapest, 1996. Auch als CIS-Report 96-7 verfügbar.

[É. Laporte 2005a] Éric Laporte. *In memoriam Maurice Gross.* Archives of Control Sciences, Bd. 15, Nr. 3, Seiten 257–278, 2005.

[É. Laporte 2005b] Éric Laporte. *Lexicon management and standard formats.* Archives of Control Sciences, Bd. 15, Nr. 3, Seiten 329–340, 2005.

[É. Laporte 2018] Éric Laporte. *Choosing features for classifying multiword expressions.* In: Manfred Sailer & Stella Markantonatou (Hgg), Multiword expressions: Insights from a multi-lingual perspective, Seiten 143–186. Language Science Press, Berlin, 2018.

[É. Laporte 2008] Éric Laporte, Takuya Nakamura & Stavroula Voyatzi. *A French Corpus Annotated for Multiword Nouns.* In: Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions, Seiten 27–30, 2008.

[L. Lemnitzer 1997] Lothar Lemnitzer. Akquisition komplexer Lexeme aus Textkorpora, Bd. 180 von *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 1997.

[J. A. Leoni de León 2008] Jorge Antonio Leoni de León. *Modèle d'analyse lexico-syntaxique des locutions espagnoles.* Dissertation, Université de Genève, Faculté des lettres, Genève, Mai 2008.

[M. Lloberes 2013] Marina Lloberes, Irene Castellón, Salvador Climent & Antoni Oliver. *Tratamiento lingüístico de multipalabras en WordNet 3.0.* In: Lirian Astrid Ciro, Lídia Gallego Balsà, Rosa Mateu Serra & Àngels Llanes (Hgg), La lingüística aplicada en la era de la globalización, Seiten 495–501. Edicions de la Universitat de Lleida, Lleida, 2013.

[E. Luis Álvarez 2019] Eva Luis Álvarez. *A Corpus Analysis of English Phraseology in Court Judgments of the European Union*, jul 2019. Advisor: Belén López Arroyo.

[G. I. Lyse 2012] Gunn Inger Lyse & Gisle Andersen. *Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text.* In: Gisle Andersen (Hg), Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian, Bd. 49 von *Studies in Corpus Linguistics*, Seiten 79–110. John Benjamins, Amsterdam/Philadelphia, 2012.

[F. Mallchok 2004] Friederike Mallchok. *Automatic Recognition of Organization Names in English Business News.* Dissertation, Ludwig-Maximilians-Universität München, München, 2004.

[C. D. Manning 2008] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, Cambridge, 2008.

[C. D. Manning 1999] Christopher D. Manning & Hinrich Schütze. Foundations of statistical natural language processing. The MIT Press, Cambridge, Massachusetts, 1999.

[G. Mapelli 2004] Giovanna Mapelli. *Locuciones del lenguaje del fútbol.* In: Domenico Antonio Cusato, Loretta Frattale, Gabriele Morelli, Pietro Taravacci & Belén Tejerina (Hgg), La memoria delle lingue: la didattica e lo studio delle lingue della Penisola iberica in Italia, Bd. 2, Seiten 171–182, Messina, 2004. Andrea Lippolis Editore.

[J. Martínez Marín 2000] Juan Martínez Marín. *Las unidades léxicas complejas en Español: Aspectos teóricos y descriptivos.* Revista de

Investigación Lingüística, Bd. 3, Nr. 2, Seiten 315–338, 2000.

[A. Martín Bosque 2008] Adelaida Martín Bosque. *¿Fare es hacer? Colocaciones en los diccionarios monolingües de aprendizaje de ELE.* In: Dolores Azorín Fernández*et al.* (Hgg), El diccionario como puente entre las lenguas y culturas del mundo: Actas del II Congreso Internacional de Lexicografía Hispánica. Biblioteca Virtual Miguel de Cervantes, 2008.

[J. Mateu 2007] Jaume Mateu & M. Teresa Espinal. *Argument structure and compositionality in idiomatic constructions.* The Linguistic Review, Bd. 24, Nr. 1, Seiten 33–59, Juni 2007.

[B. Matheny 2025] Blake Matheny, Phuong Nguyen & Minh Nguyen. *JNLP at SemEval-2025 Task 1: Multimodal Idiomaticity Representation with Large Language Models.* In: Sara Rosenthal, Aiala Rosá, Debanjan Ghosh & Marcos Zampieri (Hgg), Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Seiten 1479–1484, Vienna, Austria, Juli 2025. Association for Computational Linguistics.

[M. Mathieu-Colas 1996] Michel Mathieu-Colas. *Essai de typologie des noms composés français.* Cahiers de lexicologie, Bd. 69, Nr. 2, Seiten 71–125, 1996.

[D. Maurel 2005] Denis Maurel & Franz Guenthner. Automata and dictionaries, Bd. 6 von *Texts in Computing.* College Publications, London, 2005.

[I. Mel'čuk 1998] Igor Mel'čuk. *Collocations and Lexical Functions.* In: Anthony P. Cowie (Hg), Phraseology: Theory, Analysis, and Applications, Seiten 23–53. Oxford University Press, Oxford, 1998.

[I. Mel'čuk 2001] Igor Mel'čuk. *Fraseología y diccionario en la lingüística moderna.* In: Isabel Uzcanga Vivar, Elena Llamas Pombo & Juan Manuel Pérez Velasco (Hgg), Presencia y renovación de la lingüística francesa, Nummer 279 in Acta Salmanticensia. Estudios filológicos, Seiten 267–310. Ediciones Universidad de Salamanca, Salamanca, 2001.

[I. Mel'čuk 2006] Igor Mel'čuk. *Colocaciones en el diccionario.* In: Margarita Alonso Ramos (Hg), Diccionarios y fraseología, Bd. 3 von *Anexos de Revista de Lexicografía*, Seiten 11–44. Universidade da Coruña, Servizo de Publicacións, A Coruña, 2006.

[C. Mellado-Blanco 2008] Carmen Mellado-Blanco (Hg). Colocaciones y fraseología en los diccionarios, Bd. 44 von *Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation.* Peter Lang, Frankfurt am Main, 2008.

[V. Mititelu 2025] Verginica Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic & Ivelina Stoyanova. *Survey on Lexical Resources Focused on Multiword Expressions for the Purposes of NLP.* In: Atul Kr. Ojha, Voula Giouli, Verginica Barbu Mititelu, Mathieu Constant, Gražina Korvel, A. Seza Doğruöz & Alexandre Rademaker (Hgg), Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), Seiten 41–57, Albuquerque, New Mexico, U.S.A., Mai 2025. Association for Computational Linguistics.

[B. A. Model 2010] Benedikt A. Model. Syntagmatik im zweisprachigen wörterbuch, Bd. 137 von *Lexicographica. Series Maior.* De Gruyter, Berlin/New York, 2010.

[P. Mogorrón Huerta 2004] Pedro Mogorrón Huerta. *Los diccionarios electrónicos fraseológicos, perspectivas para la lengua y la traducción.* ELUA: Estudios de Lingüística. Universidad de Alicante, Nr. Anexo 2, Seiten 381–400, 2004.

[A. Monceaux 1995] Anne Monceaux. *Le dictionnaire des mots simples anglais: mots nouveaux et variantes orthographiques.* Rapport technique IGM 95-15, Institut Gaspard Monge, Université de Marne-la-Vallée, Marne-la-Vallée, 1995.

[J. Monti 2011] Johanna Monti, Anabela Barreiro, Annibale Elia, Federica Marano & Antonella Napoli. *Taking on new challenges in multi-word unit processing for machine translation.* In: Felipe Sánchez-Martínez & Juan Antonio

Pérez-Ortiz (Hgg), Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, Seiten 11–20, Barcelona, Spain, Januar 2011.

[E. T. Montoro del Arco 2021] Esteban T. Montoro del Arco. *La codificación de lo pluriverbal en la serie textual del gramático venezolano Baldomero Rivodó (1821–1915).* Boletín de Filología, Bd. 56, Nr. 2, Seiten 241–290, 2021.

[S. Nagel 2006] Sebastian Nagel. Formenbildung im Russischen. Formale Beschreibung und Automatisierung für das CISLEX-Wörterbuchsystem. Abschlussarbeit, Ludwig-Maximilians-Universität München, CIS, München, 2006.

[S. Nagel 2008] Sebastian Nagel. *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval.* Dissertation, Ludwig-Maximilians-Universität München, München, Juli 2008.

[A. Nandi 2007] Arnab Nandi & H. V. Jagadish. *Effective Phrase Prediction.* In: Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07), Seiten 219–230, Vienna, Austria, 2007. ACM.

[I. Nica 2004] Iulia Nica. *El conocimiento lingüístico en la desambiguación semántica automática.* Dissertation, Universitat de Barcelona, Barcelona, 2004.

[G. Nunberg 1994] Geoffrey Nunberg, Ivan A. Sag & Thomas Wasow. *Idioms.* Language, Bd. 70, Nr. 3, Seiten 491–538, September 1994.

[P. Ortega 2024] Pablo Ortega, Jordi Luque, Luis Lamiable, Rodrigo López & Richard Benjamins. *Word Sense Disambiguation in Native Spanish: A Comprehensive Lexical Evaluation Resource*, 2024.

[G. Parodi 2007] Giovanni Parodi (Hg). Working with Spanish corpora. Corpus and Discourse. Continuum, London/New York, 2007.

[C. Parra Escartín 2018] Carla Parra Escartín, Almudena Nevado Llopis & Eoghan Sánchez Martínez. *Spanish multiword expressions: Looking for a taxonomy.* In: Manfred Sailer & Stella Markantonatou (Hgg), Multiword expressions: Insights from a multi-lingual perspective, Seiten 271–323. Language Science Press, Berlin, 2018.

[S. Paumier 2016] Sébastien Paumier. *Unitex 3.1 – User Manual*, 2016.

[D. Phelps 2024] Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith & Aline Villavicencio. *Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection.* In: Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre & Alexandre Rademaker (Hgg), Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, Seiten 178–187, Torino, Italia, Mai 2024. ELRA and ICCL.

[B. Pöll 2002] Bernhard Pöll. Spanische Lexikologie - Eine Einführung. Narr Studienbücher. Gunter Narr Verlag, Tübingen, 2002.

[W. Porzig 1934] Walter Porzig. *Wesenhafte Bedeutungsbeziehungen.* Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB), Bd. 58, Seiten 70–97, 1934.

[H. Provencio Garrigós 2003] Herminia Provencio Garrigós. *Las unidades polilexemáticas de la Constitución española en los diccionarios jurídicos y generales.* In: Agustín Vera Luján, Ramón Almela Pérez, José María Jiménez Cano & Dolores Anunciación Igualada Belchí (Hgg), Homenaje al profesor Estanislao Ramón Trives, Bd. 2, Seiten 627–668. 2003.

[E. Puertas Ribés 2022] Elia Puertas Ribés. *Estudio diacrónico de locuciones nominales y adjetivales españolas en diferentes tradiciones discursivas (siglos XVI a XX).* Dissertation, Universitat Jaume I, Castellón, España, 2022. Director:

José Luis Blas Arroyo; Codirector: Mónica Velando Casanova.

[C. Ramisch 2015] Carlos Ramisch. Multiword Expressions Acquisition: A Generic and Open Framework, Bd. XIV von *Theory and Applications of Natural Language Processing*. Springer, Cham, 2015.

[C. Ramisch 2023] Carlos Ramisch. *Multiword Expressions in Computational Linguistics*. Dissertation, Aix-Marseille Université (AMU), Marseille, France, 2023. Habilitation à diriger des recherches (HDR).

[C. Ramisch 2012] Carlos Ramisch, Vitor De Araujo & Aline Villavicencio. *A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions*. In: Jackie C. K. Cheung, Jun Hatori, Carlos Henriquez & Ann Irvine (Hgg), Proceedings of ACL 2012 Student Research Workshop, Seiten 1–6, Jeju Island, Korea, Juli 2012. Association for Computational Linguistics.

[C. Ramisch 2020] Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh & Hongzhi Xu. *Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions*. In: Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova & Agata Savary (Hgg), Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Seiten 107–118, Online, Dezember 2020. Association for Computational Linguistics.

[C. Ramisch 2010] Carlos Ramisch, Aline Villavicencio & Christian Boitet. *mwetoolkit: a Framework for Multiword Expression Identification*. In: Proceedings of the Seventh

International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010. European Language Resources Association (ELRA).

[J. N. A. Ranaivoarison 2015] Joro Ny Aina Ranaivoarison. *The Malagasy Language in the Digital Age: Challenges and Perspectives.* In: Proceedings of the 7th Language and Technology Conference (LTC 2015): Human Language Technologies as a Challenge for Computer Science and Linguistics, Seiten 299–303, Poznań, Poland, November 2015.

[E. Ranchhod 1999] Ehsabete Ranchhod, Cristina Mota & Jorge Baptista. *A Computational Lexicon of Portuguese for Automatic Text Parsing.* In: SIGLEX99: Standardizing Lexical Resources, Seiten 74–80, 1999.

[Real Academia Española 1726 1739] Real Academia Española. Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar, los proverbios o refranes, y otras cosas convenientes al uso de la lengua. Francisco del Hierro, Madrid, 1726–1739.

[Real Academia Española 2014] Real Academia Española. Diccionario de la Lengua Española. Espasa Libros, Madrid, 23. Auflage, 2014.

[Real Academia Española 2025a] Real Academia Española. *Corpus del Español del Siglo XXI (CORPES XXI), versión 1.3.* Online database, 2025.

[Real Academia Española 2025b] Real Academia Española & Asociación de Academias de la Lengua Española. Nueva gramática de la lengua española. edición revisada y ampliada. Espasa Libros, Madrid, 2025.

[A. Ricós Vidal 2008] Amparo Ricós Vidal. *De locuciones coordinadas a sintagmas complejos. A propósito de a diestro y siniestro, a tuerto o a derecho, a tontas y a locas.* In: Inés Olza Moreno, Manuel Casado Velarde & Ramón González Ruiz (Hgg), Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL), Seiten

707–717. Servicio de Publicaciones de la Universidad de Navarra, Pamplona, 2008.

[P. Saenger 1997] Paul Saenger. Space between words: The origins of silent reading. Figurae: Reading Medieval Culture. Stanford University Press, Stanford, CA, 1997.

[I. A. Sag 2002] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. *Multiword Expressions: A Pain in the Neck for NLP*. In: Alexander Gelbukh (Hg), Computational Linguistics and Intelligent Text Processing (CICLing 2002), Bd. 2276 von *Lecture Notes in Computer Science*, Seiten 1–15, Berlin/Heidelberg, 2002. Springer.

[J. Sastre Alaiz 2007] Judith Sastre Alaiz. Dictionnaire électronique de catalan coordonné avec le français: le module NooJ. Abschlussarbeit, Universitat Autònoma de Barcelona, Facultat de Filosofia i Lletres, Barcelona, 2007.

[A. Savary 2005] Agata Savary. *A Formalism for the Computational Morphology of Multi-Word Units*. Archives of Control Sciences, Bd. 15, Nr. 3, Seiten 437–449, 2005.

[A. Savary 2009] Agata Savary. *Multiflex: A Multilingual Finite-State Tool for Multi-Word Units*. In: Sebastian Maneth (Hg), Implementation and Application of Automata (CIAA 2009), Bd. 5642 von *Lecture Notes in Computer Science*, Seiten 237–240, Berlin/Heidelberg, 2009. Springer.

[A. Savary 2019] Agata Savary, Silvio Ricardo Cordeiro & Carlos Ramisch. *Without lexicons, multiword expressions identification will never fly: a position statement*. In: Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (Hgg), Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Seiten 79–91, Florence, Italy, August 2019. Association for Computational Linguistics.

[A. Savary 2017] Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie

Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.* In: Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (Hgg), Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Seiten 31–47, Valencia, Spain, April 2017. Association for Computational Linguistics.

[F. Schmidt 2001] Friederike Schmidt. A Comparison of different approaches to Multi-Word Term Acquisition. Abschlussarbeit, Ludwig-Maximilians-Universität München, CIS, München, April 2001.

[F. Segond 1996] Frédérique Segond, Giuseppe Valetto & Elisabeth Breidt. *Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX.* In: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, Seiten 1036–1040, Copenhagen, Denmark, August 1996.

[V. Seretan 2011] Violeta Seretan. Syntax-based collocation extraction, Bd. 44 von *Text, Speech and Language Technology.* Springer, Dordrecht, 2011.

[V. Seretan 2004] Violeta Seretan, Luka Nerima & Eric Wehrli. *A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora.* In: Geoffrey Williams & Sandra Vessier (Hgg), Proceedings of the 11th EURALEX International Congress, Seiten 755–766, Lorient, France, jul 2004. Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

[D. Siepmann 2006] Dirk Siepmann. *Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects.* International Journal of Lexicography, Bd. 19, Nr. 1, Seiten 1–39, 2006.

[A. Siyanova-Chanturia 2017] Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan & Walter J. B. van Heuven. *Representation and processing of multi-word expressions*

in the brain. Brain and Language, Bd. 175, Seiten 111–122, 2017.

[I. Sosa Mayor 2006] Igor Sosa Mayor. Routineformeln im Spanischen und im Deutschen: eine pragmalinguistische kontrastive Analyse. Praesens Verlag, Wien, 2006.

[D. Stein 2012] Daniel Stein. *Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes*. In: Eric Atwell, Claire Brierley & Majdi Sawalha (Hgg), Proceedings of the LREC 2012 Workshop on Language Resources and Evaluation for Religious Texts (LRE-Rel), Seiten 88–94, Istanbul, Turkey, Mai 2012. European Language Resources Association (ELRA).

[A. Stojić 2012] Aneta Stojić & Nataša Košuta. *Zur Abgrenzung von Mehrwortverbindungen*. Zagreber germanistische Beiträge, Bd. 21, Nr. 1, Seiten 359–373, 2012.

[J. Tanner 2023] Joshua Tanner & Jacob Hoffman. *MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation*. In: Houda Bouamor, Juan Pino & Kalika Bali (Hgg), Findings of the Association for Computational Linguistics: EMNLP 2023, Seiten 181–193, Singapore, Dezember 2023. Association for Computational Linguistics.

[C. Timm 2010] Christian Timm. Europäischer Strukturalismus in der spanischen Grammatikographie, Bd. 516 von *Tübinger Beiträge zur Linguistik*. Narr Francke Attempto, Tübingen, 2010.

[T. Utsuro 2007] Takehito Utsuro, Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi & Satoshi Sato. *Learning Dependency Relations of Japanese Compound Functional Expressions*. In: Nicole Gregoire, Stefan Evert & Su Nam Kim (Hgg), Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Seiten 65–72, Prague, Czech Republic, Juni 2007. Association for Computational Linguistics.

[T. Van de Cruys 2007] Tim Van de Cruys & Begoña Villada Moirón. *Semantics-based Multiword Expression Extraction.* In: Nicole Grégoire, Stefan Evert & Su Nam Kim (Hgg), Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Seiten 25–32, Prague, Czech Republic, Juni 2007. Association for Computational Linguistics.

[S. Venkatapathy 2005] Sriram Venkatapathy & Aravind Joshi. *Measuring the Relative Compositionality of Verb-Noun (V-N) Collocations by Integrating Features.* In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Seiten 899–906, Vancouver, British Columbia, Canada, Oktober 2005. Association for Computational Linguistics.

[S. Vietri 2008] Simonetta Vietri. Dizionari elettronici e grammatiche a stati finiti. metodi di analisi formale della lingua italiana. Lessici e combinatorie. Plectica, Salerno, 2008.

[M. R. Vilhena Costa 2001] Maria Rute Vilhena Costa. *Pressupostos Teóricos e Metodológicos Para a Extracção Automática de Unidades Terminológicas Multilexémicas.* Dissertation, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa, Lisboa, August 2001.

[M. B. Villada Moirón 2005] María Begoña Villada Moirón. *Data-driven identification of fixed expressions and their modifiability.* Dissertation, Rijksuniversiteit Groningen, Groningen, 2005.

[J. E. Weeds 2003] Julie Elizabeth Weeds. *Measures and Applications of Lexical Distributional Similarity.* Dissertation, University of Sussex, Brighton, UK, 2003.

[F. Zagatti 2022] Fernando Zagatti, Paulo Augusto de Lima Medeiros, Esther da Cunha Soares, Lucas Nildaimon dos Santos Silva, Carlos Ramisch & Livy Real. *mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering.* In: Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) at LREC 2022, Seiten 112–117,

Marseille, France, Juni 2022. European Language Resources Association.

[W. Zhang 2008] Wen Zhang, Taketoshi Yoshida & Xijin Tang. *Text classification based on Multi-Word with Support Vector Machine*. Knowledge-Based Systems, Bd. 21, Nr. 8, Seiten 879–886, 2008.

[Z. Zhang 2008] Ziqi Zhang, José Iria, Christopher Brewster & Fabio Ciravegna. *A Comparative Evaluation of Term Recognition Algorithms*. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Seiten 2108–2111, Marrakech, Morocco, Mai 2008. European Language Resources Association (ELRA).

[J. Zhou 2024] Jianing Zhou, Ziheng Zeng, Hongyu Gong & Suma Bhat. *Enhancing Language Models with Idiomatic Reasoning*. In: Proceedings of the First Conference on Language Modeling (COLM 2024), Philadelphia, PA, USA, Oktober 2024.

[A. Zuluaga 1980] Alberto Zuluaga. Introducción al estudio de las expresiones fijas, Bd. 10 von *Studia Romanica et Linguistica*. Peter D. Lang, Frankfurt a. M.; Bern; Cirencester/U.K., 1980.

[M. T. Zurdo 2007] María Teresa Zurdo. *Phraseologie des Spanischen*. In: Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (Hgg), Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung, Seiten 703–713. Walter de Gruyter, Berlin/New York, 2007.