Zur Rolle der Flexionsmorphologie in der automatischen Klassifikation deutschsprachiger Textdokumente

Inaugural-Dissertation

zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Daniel Benedikt Claeser

aus

Dormagen

Bonn, 2022

Gedruckt mit der Genehmigung der Philosophischen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Kristian Berg
(Vorsitzender)

Prof. Dr. Claudia Wich-Reif
(Betreuerin und Gutachterin)

Prof. Dr. Ulrich Schade
(Gutachter)

Prof. Dr. Thomas Klein
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 18. November 2021

Für meine Eltern

iuxta fidem defuncti sunt omnes isti
non acceptis repromissionibus
sed a longe eas aspicientes
et salutantes
et confitentes
quia peregrini et hospites sunt supra terram.

Hebr 11, 13

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Bei der automatischen Klassifikation von natürlichsprachlichen Texten handelt es sich um eine zentrale und etablierte Disziplin der Computerlinguistik. Sie unterteilt sich in zahlreiche spezialisierte Teilgebiete – Ziel der Klassifikation kann die Erkennung etwa von Thema, Sprache, Autor, Polarität und generell jede beliebige Unterteilung von Dokumenten in im Voraus festgelegte Kategorien sein. Klassifikation kann dabei sowohl Vorverarbeitungsstufe für Dokumente sein, die einer auf ihrem Ergebnis basierenden weiteren Verarbeitung zugeführt werden sollen, als auch Endziel der automatischen Verarbeitung.

Traditionelle Klassifikationsverfahren verwenden zur Lösung ihrer Aufgabe in der Regel das sogenannte Bag-of-Words-Modell beziehungsweise dessen Weiterentwicklungen. Hierbei wird der Inhalt eines Textes vereinfachend als durch eine Anzahl ausgewählter Wortformen (im Folgenden auch Merkmale genannt) und deren Häufigkeiten repräsentiert aufgefasst. Diese Wortformen können zu übergeordneten Einheiten, sogenannten N-Grammen, zusammengestellt und in unverarbeiteter, das heißt im Deutschen in flektierter Form, oder lemmatisiert oder anderweitig vorverarbeitet verwendet werden. Die Häufigkeiten dieser Merkmale auf Dokument- wie Korpusebene bilden mit zu erlernenden Gewichtungen kombiniert die Eingabe für eine Klassifikationsfunktion. Dieses Prinzip dominierte die Entwicklung von Klassifikationsverfahren bis zum Beginn der 2010er-Jahre.

Die automatische Klassifikation von Texten erfuhr in den Jahren seit 2013 mit dem Aufkommen einer neuen Generation künstlicher neuronaler Netze eine sprunghafte Leistungssteigerung. Dieser Durchbruch basierte unter anderem auf stetig erhöhten Rechenkapazitäten und der Entwicklung von Spezialhardware. Diese technischen Fortschritte ermöglichten effiziente Umsetzungen bereits länger bekannter Techniken wie etwa *Convolutional Neural Networks (CNN)* und *Long Short-Term Memory (LSTM)*. Zusätzliche, neuentwickelte Werkzeuge wie *Attention* oder *Transformer* unterstützen diese Leistungssteigerungen erheblich.

Nicht zuletzt jedoch beruht die Dominanz verschiedener Disziplinen der Computerlinguistik durch neuronale Verfahren auf der Entwicklung des Embedding-Prinzips. Ein Embedding repräsentiert Wörter durch Vektoren, deren Belegungen auf riesigen Korpora trainiert wurden. Ein solcher Embedding-Vektor repräsentiert in Form latenter Variablen das gesamte im Eingabekorpus über dieses Wort enthaltene semantische und morphosyntaktische Wissen. Als Eingabeschicht für einen auf eine bestimmte Aufgabe trainierten neuronalen Klassifikator ermöglicht er die Wiederverwendung der in einem allgemeineren, größeren Korpus aufgefundenen Informationen über dieses Wort unter aufgabenspezifischer Gewichtung relevanter Aspekte. Diese Hinzuziehung und Interpretation vortrainierten sprachlichen Wissens ermöglicht es, die durch die Größe des Trainingskorpus vorgegebenen Leistungsgrenzen zu verschieben. Die neue Generation künstlicher neuronaler Netze hat in Verbindung mit dem Embeddingverfahren die konventionellen, frequenzbasierten Verfahren praktisch vollständig verdrängt und dominiert den aktuellen Literatur- und Konferenzbetrieb (ACL, COLING, EMNLP, AAAI).

Beiden Gruppen von Klassifikationsverfahren, traditionellen wie neuronalen, ist gemein, dass sie letztlich auf Korrelationen von Merkmalen, das heißt Wortformen, und Zielklassen beruhen: Aus diesen entsprechend trainierten Korrelationen wird mit unterschiedlich starken Gewichten auf die Zugehörigkeit eines unbekannten Dokuments zu einer oder mehreren Zielkategorien geschlossen. Basis sämtlicher Verfahren zur Merkmalsge-

winnung und des Klassifikationsvorgangs an sich bleibt also das Wort in seinen verschiedenen Erscheinungsformen. Die Zusammenhänge zwischen Lexemen und deren Repräsentationen im Text sowie zwischen Lexem und Klasse müssen vom Klassifikator aus einem annotierten Trainingskorpus erlernt werden. Flektierende Sprachen bedienen sich einer Reihe von Mechanismen wie Affigierung und Umlautung, um semantische oder morphosyntaktische Informationen in einer Wortform zu kodieren. Die Kodierung von Informationen wie Numerus und Kasus direkt am Wort statt über die syntaktische Position oder separate Partikeln führt zur Realisierung eines Lexems in einer Reihe von Einzelformen. Diese erscheinen in einem Text als unterschiedliche Zeichenketten, die für einen Klassifikator nicht ohne Weiteres als demselben Lexem zugehörig erkennbar sind. Umgekehrt können durch morphologische Veränderungen flektierte Formen mehrerer Lexeme in einer Zeichenkette als Homographen zusammenfallen. Beide Phänomene haben das Potenzial, den Lern- und Klassifikationsvorgang eines Bag-of-Words-Klassifikators zu beeinträchtigen.

Die vorliegende Studie untersucht als erste dieser Art spezifisch den Einfluss der Flexionsmorphologie der deutschen Sprache auf Training und Betrieb eines automatischen Klassifikationssystems auf deutschsprachigen Texten. Gegenstand der Studie sind hingegen nicht die Nachbarfelder Kompositions- und Derivationsmorphologie: Beide Vorgänge erzeugen, unter Umständen wortklassenübergreifend, neue Wörter aus existierenden Lexemen, die sodann den Regeln der Flexionsmorphologie unterworfen sind. Komposition und Derivation werden in dieser Studie als der Flexion vorgelagerte Prozesse der Entstehung von potenziellen Merkmalen betrachtet. Sie werden daher als separater Untersuchungsgegenstand angesehen und in dieser Studie nicht weiter besprochen.

Auch wenn es sich bei der vorliegenden Studie um die erste systematische Untersuchung flexionsmorphologischer Phänomene der deutschen Sprache speziell in Bezug auf ihre Wirkung in der Textklassifikation handeln dürfte, ist eine Reihe der verwendeten Konzepte jeweils separat bereits früh in der Computerlinguistik thematisiert worden: Bereits

Weber (1973) identifiziert Homographie als Phänomen von Interesse in der Computerlinguistik und beschreibt ein algorithmisches Verfahren zur automatischen Generierung eines Homographiewörterbuchs. Fabricz (1986) thematisiert möglicherweise erstmals in multilingualer Perspektive die Homographie von Partikeln wie Adverbien im Zusammenhang mit maschineller Übersetzung. Krovetz (1997) unterscheidet erstmals klar zwischen Homonymie und Polysemie und benennt beide als potentielle Problemquellen im Information Retrieval (und somit indirekt in der automatischen Textklassifikation). Klavans and Kan (1998) identifizieren Verben als relevante Informationsquelle für die Kategorisierung englischsprachiger Nachrichtentexte (hier als semantische Thematik ohne Bezug zu (Flexions-)Morphologie aufgefasst). Lezius et al. (1998) dokumentieren für ihr integriertes System *Morphy* für deutschsprachiges POS-Tagging und Lemmatisierung domänenintern eine Accuracy von 96% für ein kleines und 85% für ein erweitertes Tagset sowie eine Accuracy von 99,3% für Lemmata, die im Lexikon gefunden oder über den POS-Tagger disambiguiert werden können. ten Hacken and Bopp (1998) identifizieren trennbare Verbpartikeln als potentielles Problem in der Textanalyse im Sinne dieser Untersuchung und schlagen eine regelbasierte Auflösungsstrategie vor. Pachunke et al. (1992) präsentieren ein erstes Wörterbuch zur morphologischen Segmentierung für Deutsch mit etwa 11.000 Einträgen. Powers (1998) diskutiert Zipfs Gesetz im Hinblick auf das zunehmende Auftreten von seltenen Termen und Hapaxlegomena (also auch flektierter Wortformen) mit wachsender Korpusgröße.

Die Arbeit ist in fünf Kapitel unterteilt. Das auf die Einleitung folgende Kapitel 2 bietet einen Überblick über die wesentlichen Konzepte der automatischen Textklassifikation. Da es sich um ein weites Feld der Computerlinguistik handelt, müssen sich die Ausführungen auf grundlegende Methodik und Terminologie beschränken; nur die im empirischen Teil der Arbeit verwendeten Klassifikationsverfahren werden detailliert vorgestellt. Kapitel 3 beschreibt die in dieser Arbeit verwendeten linguistischen Ressourcen, bestehend aus Korpora und Wörterbuch, und Softwarepakete, das heißt, Klassifikationssoftware

und vortrainierte Embeddings. Kapitel 4 stellt die flexionsmorphologischen Phänomene Deklination und Konjugation sowie deren Einfluss auf das Entstehen von Homographien vor. Das Phänomen trennbarer Verbpartikeln wird dabei als Erscheinung der Konjugation aufgefasst und diskutiert. Den Abschluss des Kapitels bildet eine Reihe von Formalisierungen zum Einfluss dieser Phänomene auf Merkmalsauswahl und Klassifikation. Kapitel 5 dokumentiert eine Reihe exemplarisch auf einem Korpus von Nachrichtentexten durchgeführter Experimente. Diese untergliedern sich in die Gruppen *Experimente zur Merkmalsauswahl* und *Klassifikationsexperimente*. Letztere wiederum unterteilen sich nach Art des genutzten Klassifikators, wobei konventionelle Klassifikatoren, aber auch ein künstliches neuronales Netz, das Embeddings nutzt, betrachtet werden. Kapitel 6 fasst die gewonnenen Erkenntnisse zusammen und gibt einen Überblick über offene Forschungsfragen sowie einen Ausblick auf zukünftige Entwicklungen.

2. Theorie der Klassifikation

Die folgende Einführung grundlegender Konzepte und Terminologie der Textklassifikation folgt, soweit nicht gesondert ausgewiesen, der Darstellung in Schütze et al. (2008).

Ein *Textklassifikator* ist zunächst eine Funktion γ, die eine Menge von Dokumenten $\mathbb{X}$ auf eine Menge von Klassen $\mathbb{C}$ abbildet:

$$\gamma : \mathbb{X} \rightarrow \mathbb{C} \tag{2.1}$$

Das bedeutet, dass sie einem Dokument d aus der Menge $\mathbb{X}$ eine oder mehrere Klassen aus einer Menge $\mathbb{C}$ zuweist. Besteht $\mathbb{C}$ nur aus zwei Klassen, spricht man von binärer Klassifikation. Hierunter fällt auch die Frage, ob ein Dokument zu einer Klasse gehört oder nicht. Enthält $\mathbb{C}$ mehr als zwei Klassen, spricht man von *Multiklassenklassifikation* (englisch *multi-class classification*). Kann einem Dokument mehr als eine Klasse zugewiesen werden, etwa in Form einer Menge möglicher Klassen mit absteigendem Anteil oder absteigender Wahrscheinlichkeit, spricht man von *Multi-Label-Klassifikation*. Im weiteren Sinne bezeichnet der Begriff *Klassifikator* auch eine Klassifikationen durchführende Software oder -bibliothek. In dieser Arbeit wird, soweit nicht anders gekennzeichnet, diese erweiterte Bedeutung des Begriffs verwendet. In der wissenschaftlichen Praxis werden Klassifikatoren unter Verwendung einer Reihe etablierter Softwarebibliotheken generiert,

parametrisiert und evaluiert. Diese Schritte und eine Auswahl in dieser Untersuchung verwendeter Klassifikationsverfahren beschreibt das vorliegende Kapitel. Unterkapitel 2.1 erklärt dazu die wesentlichen Komponenten eines Klassifikationssystems und grundlegendes Terminologisches, Unterkapitel 2.2 widmet sich der Auswahl von Klassifikationsmerkmalen, Unterkapitel 2.3 den gängigen Metriken zur Evaluation eines Klassifikationssystems. Das letzte Unterkapitel, 2.4, stellt die in dieser Untersuchung verwendeten Klassifikatoren, unterteilt in lineare, nichtlineare und neuronale Klassifikationsverfahren, vor.

2.1. Aufbau eines Klassifikationssystems

Ein Klassifikator, der mit den Mitteln des Machine Learnings entwickelt wird, erlernt aus einer Menge von Beispieldokumenten die Parameter einer Funktion, mit der die Klassenzugehörigkeit neuer, unbekannter Dokumente postuliert werden kann. Diese Menge von Beispieldokumenten wird *Trainingskorpus* genannt, die Bezeichnungen ihrer Klassenzugehörigkeiten werden *Labels* genannt. Der Vorgang der Zuweisung dieser Labels zu Dokumenten durch einen menschlichen *Annotator* oder *Trainer* zur Erstellung von Trainingsmaterial wird als *Annotation* oder *Labeling* bezeichnet. Der Vorgang des Erlernens der Parameter der Funktion durch den Klassifikator wird *Training* genannt. Die Menge der Klassen, die Anzahl der einem Dokument zuzuweisenden Klassen und die Größe des Korpus nach Tokens und Dokumenten sind anwendungsspezifisch und grundsätzlich beliebig groß. Die Genauigkeit, mit der die Trainingsdaten abgebildet oder modelliert werden können, wird zunächst nach einer auszuwählenden Metrik auf einem weiteren Korpus, dem *Evaluations-* oder *Development Corpus*, ermittelt. Die Parameter der Funktion werden iterativ über den Abgleich mit diesem Teilkorpus optimiert. Nach einer vorgegebenen Anzahl von Iterationen oder einem Abbruchkriterium wie einem bestimmten Konvergenzgrad endet das Training. Es erfolgt nun ein Test auf einem dritten Teilkorpus,

dem eigentlichen *Testkorpus*, das bisher noch nicht gesehene Daten enthält. Die Metriken für die Beurteilung der Leistungsfähigkeit eines Klassifikators werden in Unterkapitel 2.3 behandelt. Unabhängig von der gewählten Metrik sind die ermittelten Werte auf dem zum Test verwendeten Korpusausschnitt allein nicht sehr aussagekräftig: Sie besagen grundsätzlich nur, dass das auf genau diesen Trainingsdaten erlernte Modell auf genau diesen Testdaten Ergebnisse der ermittelten Qualität ermöglicht, aber nicht, dass diese Ergebnisse allgemein auf *Testdaten dieser Art* mit Parametern erreicht werden können, die auf *Trainingsdaten dieser Art* erlernt wurden. Zu diesem Zweck werden Training, Evaluation und Test mehrfach wiederholt, wobei die drei Teilmengen der ingesamt zur Verfügung stehenden Beobachtungen jedesmal neu gemischt werden. Als bestes Modell gilt jenes mit den über alle Teilkombinationen durchschnittlich besten Ergebnissen. Dieses Vorgehen zur Absicherung der Ergebnisse über die Varianz einzelner Durchläufe wird *Kreuzvalidierung* genannt.

Die genannten Teilkorpora können in unverarbeiteter oder vorverarbeiteter Form vorliegen, wobei sie in analoger Weise verarbeitet sein müssen. Dies bedeutet vor allem, dass ein Klassifikator, der auf Dokumenten in einem bestimmten Verarbeitungszustand trainiert wurde, Dokumente nur im gleichen Verarbeitungszustand erfolgreich klassifizieren kann. Traditionelle Vorverarbeitungsschritte sind etwa Lemmatisierung und Stemming (Letzteres eher im englischsprachigen Bereich).

2.1.1. Zipfs Gesetz

Zipfs Gesetz (Zipf (1949)), angewendet auf die Verteilung von Wörtern in natürlichsprachlichen Texten, besagt, dass die Wahrscheinlichkeit des Auftretens eines Wortes umgekehrt proportional zu seinem Rang in einer Häufigkeitenliste eines beliebigen Korpus ist, so dass gilt

$$p(n) \approx \frac{1}{n * ln(1,78N)} \tag{2.2}$$

wobei n den Rangplatz des Wortes und N die Größe des Vokabulars bezeichnet.

Aus der mit zunehmendem Rangplatz geringeren Wahrscheinlichkeit des Auftretens neuer Wortformen und damit Informationen steht zu vermuten, dass die Lernrate eines Klassifikators mit zunehmender Trainingsmenge immer geringer wird. Zipfs Gesetz ist daher für Konzeption und Training von Klassifikationssystemen von zentraler Bedeutung, setzt es doch offensichtlich der Möglichkeit, einen Klassifikator durch das Zurverfügungstellen einer immer größeren Trainingsmenge beliebig zu verbessern, enge Grenzen.

2.1.2. Das Vektorraummodell

Etablierte Klassifikationsverfahren arbeiten mit reellwertigen numerischen Repräsentationen der vorgelegten Dokumente in Form von Vektoren fixer Dimension anstelle der Originaldokumente in Textform. Während die Besetzungen dieser Vektoren für neuronale Klassifikationsverfahren keinen Beschränkungen unterliegen, sind sie im Falle konservativer, nichtneuronaler Klassifikationsverfahren auf positiv semidefinite Werte beschränkt. Bei der Mehrzahl der etablierten Klassifikationsverfahren erlernt der Klassifikator einen für jede Klasse spezifischen Vektor mit Gewichten oder Koeffizienten. Dessen Skalarprodukt mit den Werten eines Dokumentvektors bildet in der einen oder anderen Form, das heißt mit oder ohne Transformation, die Basis für die Bildung des *Scores* einer Klasse.

In der einfachsten Form kann die Überführung eines Dokumentes in einen reellwertigen Vektor in Form einfacher Auszählung der absoluten Häufigkeiten einer jeden vorkommenden Wortform erfolgen. Dieser wenig aussagekräftige Wert sollte im mindesten Fall

in irgendeiner Form normalisiert werden, etwa auf die Länge des Dokumentes, womit der Vektor, der durch die Auszählung der Wortformen erzeugt wurde, auf die Länge 1 gebracht wird.

Ein aussagekräftiger normalisiertes Standardmaß ist die *TF-IDF* (für *Termfrequenz-Inverse Dokumentenfrequenz*, englisch *term frequency-inverse document frequency)*. Sie formuliert ein Merkmal m auf Basis der absoluten Häufigkeit eines Terms in einem Dokument unter Berücksichtigung der generellen Häufigkeit des Terms im gesamten Trainingskorpus:

$$tf - idf_{t,d} = tf_{t,d} * idf_t \tag{2.3}$$

Die *Termfrequenz* (englisch *term frequency*) *TF* ist definiert als absolute Häufigkeit eines Terms im Dokument, die *Inverse Dokumentenfrequenz IDF* als

$$idf_t = log \frac{N}{df_t} \tag{2.4}$$

wobei df_t die Dokumentenfrequenz (englisch *document frequency*) bezeichnet, das heißt, die Anzahl der Dokumente, in denen der Term vorkommt. Die Anzahl aller Trainingsdokumente insgesamt N wird dividiert durch die Anzahl der den Term enthaltenden Dokumente, so dass die IDF umso höher ausfällt, je weniger Dokumente das Wort enthalten. Mit diesem Verfahren werden im Korpus Terme, die nur in sehr wenigen Dokumenten auftreten, höher gewichtet als häufige Terme wie Partikeln aller Art, so dass Unterschiede zwischen Dokumenten in dieser Hinsicht im Klassifikationsverfahren größere

Auswirkungen haben. In vielen Dokumenten jeweils genau einmal auftretende, vergleichsweise seltene Terme erzeugen so geringere Werte als ein Term, der insgesamt genau so häufig auftritt, aber nur in wenigen Dokumenten. Die Logarithmierung des Quotienten resultiert in einer zunehmenden Stauchung der TF-IDF mit abnehmender Dokumentenfrequenz, so dass immer seltenere Terme nicht immer stärker „belohnt“ werden, was unerwünschterweise auch Ausreißertermen und Hapaxlegomena zugute käme.

Über unlemmatisierte oder lemmatisierte Terme hinaus können auch N-Gramme beliebiger Länge auf diese Weise transformiert und somit berücksichtigt werden. Auch weitere sprachliche Muster wie etwa *Part-of-speech-tags* können analog in numerische Merkmale überführt werden.

Die Vektorisierung von natürlichsprachlichen Texten ermöglicht in einem ersten Anwendungsschritt einen Vergleich dieser Dokumente mit trigonometrischen Methoden. Bei der *Kosinusähnlichkeit* zweier Vektoren $\vec{u}$ und $\vec{v}$ gleicher Dimension, definiert als

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2}\sqrt{\sum_{i=1}^{n} (v_i)^2}} \tag{2.5}$$

resultierend aus

$$\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta(\vec{u}, \vec{v}) \tag{2.6}$$

handelt es sich um ein bereits frühzeitig im *Information Retrieval* verwendetes, verbreitetes Ähnlichkeitsmaß. Es handelt sich hierbei um das Skalarprodukt dieser beiden Vektoren, normalisiert auf das Produkt ihrer euklidischen Längen. Durch die Normalisierung wird ein Wertebereich zwischen 0 und 1 garantiert. Hierbei ergibt sich eine

Kosinusähnlichkeit von 0 bei zwei Vektoren ohne jede Ähnlichkeit, was einem Winkel von 90 Grad entspricht, und eine Kosinusähnlichkeit von 1 bei zwei Vektoren, die exakt gleich sind, also mit einem Winkel von 0 Grad aufeinanderliegen. Eine theoretisch mögliche Kosinusähnlichkeit von -1 bei 180 Grad kann sich bei Vektoren, die aus Frequenzen von Merkmalen in Dokumenten gebildet werden, nicht ergeben, da hierzu Werte kleiner als Null erforderlich wären, was bei einer minimal möglichen Dokumentenfrequenz von 0 nicht der Fall sein kann.

Die Nutzbarkeit der Kosinusähnlichkeit zum direkten Vergleich zweier Dokumentvektoren ist Grundlage zweier in dieser Arbeit verwendeter Klassifikationsverfahren (s. Unterkapitel 2.4). Das Skalarprodukt eines Dokumentenvektors mit einem Koeffizienten- oder Gewichtungsvektor wiederum ist ein wiederkehrender Baustein einer Reihe weiterer Klassifikationsverfahren.

2.2. Merkmalsauswahl am Beispiel des Chi-Quadrat-Tests (χ^2)

Vor der Auswahl eines geeigneten Klassifikationsalgorithmus stellt sich die Frage nach der Bestimmung der zur Klassifikation heranzuziehenden Merkmale: Im einfachsten Fall könnte schlicht das gesamte im Trainingskorpus verfügbare Vokabular zur Klassifikation genutzt werden. Dieses Vorgehen ist aus zweierlei Gründen nicht praktikabel: Nach Zipfs Gesetz tritt ein Großteil der Wortformen eines Korpus nur wenige Male auf und trägt somit nur wenig Informationen zur Klassifikation bei. Der Aufwand für das Erlernen der Gewichtungen zu diesen Merkmalen und ihre Kombination wächst jedoch exponentiell für zunehmend geringeren Informationsgewinn. Der weitaus gravierendere Nachteil dieser Vorgehensweise besteht jedoch in dem Umstand, dass bei Verwendung sämtlicher verfügbarer Merkmale das resultierende Modell zu speziell auf die vorgelegten Trainings-

und Evaluationsdokumente zugeschnitten ist. Das Vorhandensein oder Nichtvorhandensein von Merkmalen, die nur sehr spezielle Aspekte der zu beschreibenden Kategorien repräsentieren, in bisher ungesehenen Dokumenten reduziert die *Generalisierungsfähigkeit* des Modells. Aus diesen Gründen ist ein möglichst kompaktes Modell, das einen Kompromiss aus möglichst hoher Abdeckung der *Varianz* der Trainingsdaten bei möglichst geringem *Bias* zugunsten der konkreten Trainingsdokumente abbilden kann, zu bevorzugen.

Die der allgemeinen Statistik entnommene χ^2-Teststatistik,

$$\chi^2 = \frac{(O_k - E_k)^2}{E_k} = \sum_{e_t \in 1,0} \sum_{e_c \in 1,0} \frac{(O_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}, \tag{2.7}$$

bemisst den Grad der Abhängigkeit zweier Variablen, hier das gemeinsame Auftreten eines Terms und einer Klasse, voneinander. Hierbei entspricht e_t dem Auftreten des Terms im Dokument und e_c der Zugehörigkeit des Dokuments zu einer Klasse (Hull et al. (1996)). Beide Variablen können die Werte 0 oder 1, alternativ *wahr* oder *falsch* annehmen. O bezeichnet ferner das beobachtete Auftreten der vier möglichen Kombinationen dieser Belegungen und E das erwartete Auftreten.

Die Abweichung zwischen erwartetem und tatsächlichem gemeinsamen Auftreten von Term und Klasse wird zur Neutralisierung des Vorzeichens und zur Verstärkung größerer Abweichungen quadriert und sodann auf den Erwartungswert normalisiert, um nicht häufige Terme per se zu belohnen. Bei dem resultierenden Wert handelt es sich mithin um ein Maß, das mit zunehmender Diskrepanz zwischen zufälliger Verteilung und tatsächlich überproportionalem Auftreten eines Terms in einer Klasse steigt. Schütze et al. (2008) merken an, dass dieses Maß unter informationstheoretischen Gesichtspunkten auf natürlichsprachliche Texte angewandt als nicht unproblematisch anzusehen sei. Resultat eines χ^2-Tests mit sämtlichen verfügbaren Termen eines Korpus ist eine absteigend sortierbare Liste, die eine Auswahl der n Terme mit den höchsten χ^2-Werten als

Merkmale zur Eingabe in den Klassifikator ermöglicht. Auch wenn die resultierenden Einzelwerte im Hinblick auf die zugrundeliegende Distribution, insbesondere die zahlreich auftretenden seltenen Terme betreffend, nicht unproblematisch zu interpretieren sind, spiegelt ihre Ordnung dennoch die Wichtigkeit der Terme in für den angestrebten Zweck hinreichendem Ausmaß wider.

Der χ^2-Test hat sich in der Praxis gegen alternative Merkmalsauswahlverfahren wie *Information Gain* und *Dokumentenfrequenz* durchgesetzt (Yang and Pedersen (1997), Rogati and Yang (2002), weniger deutlich (für kleine bis mittlere Merkmalsmengen) Forman et al. (2003)). Sämtliche quantitativ vergleichenden Arbeiten legen dabei aber das englischsprachige Reuters-Nachrichtenkorpus zugrunde (s. auch Unterkapitel 3.1). Vergleichende Arbeiten für deutschsprachige Textklassifikation scheinen nicht vorzuliegen – eine solche sprachspezifische Untersuchung erschiene jedoch angesichts des Umstandes, dass morphologische Produktivität die Anzahl und Verteilung von Wortformen in einem Korpus maßgeblich beeinflusst, sinnvoll, zumal gerade im Falle des χ^2-Tests mit der Problematik geringer Aussagekraft bei niedrigen Häufigkeiten argumentiert wird und die Häufigkeit der einzelnen Formen naturgemäß sinkt, wenn von einem Lexem aufgrund der produktiven Morphologie mehr Formen existieren. Mit dem Durchbruch neuronaler Klassifikatoren mit ihren impliziten Merkmalsauswahlverfahren kam jedoch die Arbeit an Methoden zur vorgelagerten Merkmalsauswahl de facto zum Erliegen.

2.3. Evaluation eines Klassifikationssystems

Sowohl für die graduelle Steuerung des Lernvorgangs als auch zur abschließenden Bewertung der Leistungsfähigkeit eines Klassifikationssystems sind einheitliche, aber flexibel priorisierbare Kennzahlen beziehungsweise Metriken etabliert. Es handelt sich dabei um

den sogenannten *F-Score*, der sich aus den gewichteten Teilwerten *Precision* und *Recall* zusammensetzt, sowie die seltener verwendete *Accuracy*.

Precision benennt den Anteil der einer Klasse zugeordneten Dokumente, die auch tatsächlich dieser Klasse angehören, definiert als

$$P = \frac{tp}{tp + fp} \tag{2.8}$$

und *Recall* den Anteil einer beliebigen Klasse zugehöriger Dokumente, die dieser auch tatsächlich zugeordnet wurden, definiert als

$$R = \frac{tp}{tp + fn} \tag{2.9}$$

wobei tp („true positive“) die Anzahl der korrekt zugeordneten Dokumente bezeichnet, fp („false positive“) die Anzahl der Texte, die unzutreffenderweise der Klasse zugeordnet wurden und fn („false negative“) die Anzahl der Texte, die nicht der Klasse zugeordnet wurden, obwohl dies hätte erfolgen müssen.

Der F-Score bildet ein gewichtetes harmonisches Mittel aus Precision und Recall als

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2.10}$$

wobei

$$\beta = \frac{1-\alpha}{\alpha} \tag{2.11}$$

mit $\alpha \in [0,1]$ und somit $\beta^2 \in [0,\infty]$ und bietet somit die Möglichkeit, einen Klassifikator mit einem gewünschten Schwerpunkt auf Precision oder Recall zu evaluieren.

Für eine Klassifikationsaufgabe mit mehr als zwei Klassen kann die durchschnittliche Leistungsfähigkeit über alle Klassen grundsätzlich auf zwei verschiedene Arten berechnet werden: als *Micro-average* oder als *Macro-average* von Precision oder Recall.

Die Ermittlung des Durchschnittswertes mittels Micro-averaging errechnet Precision und Recall nach den obenstehenden Gleichungen über alle Dokumente in einem Klassifikationsszenario als

$$P_{micro} = \frac{\sum_{i=1}^{|C|} tp_i}{\sum_{i=1}^{|C|} tp_i + fp_i} \tag{2.12}$$

respektive

$$R_{micro} = \frac{\sum_{i=1}^{|C|} tp_i}{\sum_{i=1}^{|C|} tp_i + fn_i} \tag{2.13}$$

und ermöglicht somit eine Gewichtung der Ergebnisse einzelner Klassen entsprechend ihrer Größe: Kleinere Klassen werden auf diese Weise weniger stark berücksichtigt als

größere. Dieses Verfahren ist das Mittel der Wahl bei der Evaluation in einem Szenario mit zwischen den Klassen ungleich verteilten Dokumenten.

Die alternative gleichstarke Gewichtung aller Klassen bei tatsächlicher oder mangels anderer Informationen vermuteter Gleichverteilung berechnet sich als

$$P_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{tp_i}{tp_i + fp_i} \tag{2.14}$$

und

$$R_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{tp_i}{tp_i + fn_i} \tag{2.15}$$

Die Werte werden also zunächst für jede Kategorie getrennt erhoben und sodann über die Kategorien ohne unterschiedliche Gewichtung gemittelt, so dass jede Kategorie einen gleich großen Anteil zum Ergebnis beiträgt.

Accuracy, die seltener verwendete Alternative zu Precision und Recall, berechnet den Klassifikationserfolg als

$$A = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.16}$$

also den Anteil der auf beide Arten korrekt getroffenen Zuordnungen am Anteil aller erfolgten Zuordnungen.

2.4. Klassifikationsverfahren

Nach der Festlegung der zu verwendenden Merkmale und ihrer quantitativen Auswahl durch ein geeignetes Verfahren wie χ^2 kann der Klassifikator trainiert und evaluiert werden. Dieser Abschnitt stellt einige verbreitete Klassifikationsverfahren vor: Die zu den linearen Klassifikatoren gehörenden *Naive Bayes*, *Logistische Regresson*, *Support Vector Machine* und *Rocchio*, den nichtlinearen *K-nächste-Nachbarn-Klassifikator* sowie die Merkmale beider Gruppen aufweisenden *Künstlichen Neuronalen Netze*.

2.4.1. Lineare Klassifikatoren

Ein linearer Klassifikator ist eine Funktion der Form

$$y = f(\vec{w} \cdot \vec{x}) = f(\sum_{i=1}^{n} w_i x_i), \tag{2.17}$$

die einer Beobachtung, etwa einem Text, einen klassenspezifischen Zielwert oder Score zuordnet. Hierbei ist $\vec{x}$ ein reellwertiger Merkmalsvektor zu einem Dokument und $\vec{w}$ ein klassenspezifischer reellwertiger Vektor mit Gewichten zu diesen Merkmalen. Bei den Merkmalen handelt es sich im Zusammenhang mit natürlichsprachlichen Dokumenten in der Regel um die numerischen Werte zu Häufigkeiten von Uni- oder N-Grammen von Wörtern wie etwa TF-IDF. Die mit einer solchen Funktion erzeugten klassenspezifischen Scores können absteigend sortiert und zur Zuordnung eines Dokuments zu einer oder mehreren Klassen verwendet werden. Gegebenenfalls kann der Wert der linearen Kombination der Merkmale über eine Schwellenwert- oder Vorzeichenfunktion f konvertiert werden.

2.4.1.1. Der Naive Bayes-Klassifikator

Nach dem Theorem von Bayes ist die Wahrscheinlichkeit eines Ereignisses A gegeben Ereignis B als

$$P(A|B) = \frac{P(B|A)P(B)}{P(A)} \tag{2.18}$$

zu berechnen, wobei $P(A)$ die A-priori-Wahrscheinlichkeit von Ereignis A bezeichnet, $P(B|A)$ die Wahrscheinlichkeit für B, wenn A eingetreten ist, und $P(B)$ die A-priori-Wahrscheinlichkeit für den Eintritt von Ereignis B.

Dieser Zusamenhang kann für die Klassifikation genutzt werden, wenn man die Zugehörigkeit des Dokuments d zur Klasse c als Ereignis A auffasst und das Auftreten des Terms t_k in Dokument d als Ereignis B (Maron and Kuhns (1960)). Es gilt nun

$$P(c|t_k) = \frac{P(t_k|c)P(t_k)}{P(c)} \tag{2.19}$$

wobei t_k den für die Klassifikation herangezogenen Term bezeichnet.

Die Funktion für die Klassifikation eines Dokuments d_c mithilfe einer Menge von k Termen lautet somit

$$P(d_c) = \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \tag{2.20}$$

In den wenigsten Fällen tragen Terme eine starke Indikation zur Entscheidung über die Zugehörigkeit zu einer Klasse bei, so dass die Multiplikation einer großen Zahl kleiner Wahrscheinlichkeiten in der klassifizierenden Software schnell zu einer Gleitkomma-Ausnahme führen kann. Ausgehend von

$$\log xy = \log x + \log y \tag{2.21}$$

ist die tatsächlich implementierte Funktion daher in der Regel

$$P(d_c) = log(\hat{P}(c)) + \sum_{1 \leq k \leq n_d} log(\hat{P}(t_k|c)) \tag{2.22}$$

Zur Bezifferung der A-priori-Wahrscheinlichkeit existieren zwei etablierte Schätzverfahren: Das *multinomiale Modell* schätzt die die A-priori-Wahrscheinlichkeit des Auftretens von Term t_k als

$$\hat{P}(t_k|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B'} \tag{2.23}$$

wobei T_{ct} die Häufigkeit des Terms in allen Texten der Klasse c ist und es sich bei B um die Gesamtgröße des Vokabulars handelt. Die Addition von 1 zu jeder Termhäufigkeit und deren Ausgleich durch B wird *La Place-Smoothing* genannt und dient der Kompensation von Häufigkeiten mit dem Wert Null.

Das *Bernoulli-* oder *binomiale Modell* schätzt die A-priori-Wahrscheinlichkeit für einen Term t_k als

$$\hat{P}(t_k|c) = \frac{N_{ct}}{N_c} \tag{2.24}$$

wobei N_{ct} die Anzahl der den Term enthaltenden Dokumente in Klasse c bezeichnet und N_c die Anzahl der Dokumente in Klasse c insgesamt.

2.4.1.2. Logistische Regression

Die Klassifikationsfunktion der *logistischen Regression* (Schütze et al. (1995)),

$$f(x) = \frac{1}{1 + e^{-(\sum_{i=1}^{n} w_i x_i)}} \tag{2.25}$$

transformiert die lineare Kombination der gewichteten Merkmale durch ihre Verwendung als negativen Exponenten einer Exponentialfunktion. Da die Gewichte der Merkmale beliebig besetzt werden können, ist der Wertebereich weder im Positiven noch Negativen begrenzt, so dass der Term beliebig groß oder klein werden kann. In Addition mit der Konstante 1 im Teiler garantiert er so einen Rückgabewert der Klassifikationsfunktion zwischen 0 und 1, wobei beide Grenzwerte niemals erreicht werden. Der Rückgabewert kann sodann als Schwellenwert zur Entscheidung für oder gegen eine Klassenzuweisung interpretiert werden.

2.4.1.3. Support Vector Machine

Die *Support Vector Machine* (Cortes and Vapnik (1995), Joachims (1998)) wählt aus einer Menge von Trainingsvektoren eine Reihe sogenannter Stützvektoren (*Support vectors*) zur Definition einer möglichst breit trennenden *Hyperebene*.

Dabei soll die Hyperebene, als das, was im dreidimensionalen Raum eine Ebene und was im zweidimensionalen Raum eine Gerade darstellt, die Vektoren, die Dokumente einer Klasse repräsentieren, möglichst gut von den Vektoren trennen, die Dokumente anderer Klassen repräsentieren.

In ihrer simpelsten Erscheinungsform unter Verwendung eines einzelnen Trainingsvektors handelt es sich um einen linearen Klassifikator der Form

$$f(x) = sgn(\vec{w}^\top \vec{x} + b) \tag{2.26}$$

wobei $\vec{w}$ einen zu erlernenden Gewichtungsvektor zu Trainingsvektor $\vec{x}$ bezeichnet und b den zu erlernenden, sogenannten *Intercept* oder *Offset* der Hyperebene.

Der Abstand eines Stützvektors $\vec{x_i}$ am Rande der Hyperebene mit der erwünschten maximalen Breite zur Entscheidungsgrenze in deren Mitte berechnet sich aus

$$r_i = y\frac{\vec{w}^\top \vec{x_i} + b}{|\vec{w}|} \tag{2.27}$$

mit $y \in \{1, -1\}$ zur Bezeichnung einer binären Klassenzugehörigkeit. Die vorgenommene Normalisierung ermöglicht die Formulierung einer Anforderung

$$y_i(\vec{w}^\top \vec{x_i} + b) \geq 1 \tag{2.28}$$

für alle Trainingsdatenpunkte, von denen einige als namensgebende Stützvektoren mit $y_i(\vec{w}^\top \vec{x_i} + b) = 1$ die Begrezung der Hyperebene positionieren.

Die zu maximierende Breite der Hyperebene beträgt

$$\rho = 2/|\vec{w}|, \tag{2.29}$$

umformbar in das Ziel der Minimierung von

$$\rho = |\vec{w}|/2 \tag{2.30}$$

Hiermit kann die SVM als Minimierungsproblem für $\vec{w}$ und b unter den Bedingungen

$$argmin(\frac{1}{2}\vec{w}^{\top}|\vec{w}|) \tag{2.31}$$

sowie

$$y_i(\vec{w}^{\top}\vec{x_i} + b) \geq 1 \tag{2.32}$$

für alle $\{(\vec{x_i}, y_i)\}$ formuliert werden.

Die endgültige Klassifikationsfunktion unter Einbindung der Nebenbedingungen in Form von Lagrange-Multiplikatoren lautet

$$f(x) = sgn(\sum_i \alpha_i y_i \vec{x_i}^\top \vec{x} + b) \tag{2.33}$$

und ist als quadratisches Optimierungsproblem zu lösen. α_i und y_i stehen für die dem Trainingsbeispiel zugehörenden Lagrange-Multiplikator respektive das Klassenlabel (-1 oder +1).

Der Umstand, dass die Klassifikationsentscheidung letztlich auf Skalarprodukten der Trainings- und Gewichtungsvektoren beruht, kann über eine Manipulation dieses Produktes durch sogenannte Kernelfunktionen genutzt werden, um komplexere, nichtlineare Merkmalskombinationen abzubilden. Die gängigsten Kernelfunktionen sind *polynomiale* Kernel der Form

$$K(\vec{x}, \vec{z}) = (1 + \vec{x}^\top \vec{z})^d \tag{2.34}$$

sowie RBF-Kernel der Form

$$K(\vec{x}, \vec{z}) = \exp(-\frac{(\vec{x} - \vec{z})^2}{2\sigma^2}) \tag{2.35}$$

Die Beschreibung der mathematischen Grundlagen der SVM erfolgte an dieser Stelle ausführlicher, da diese Klassifikatoren bis zur Einführung der aktuellen Generation neuronaler Netze zu den durchsetzungsfähigsten in wissenschaftlichen Wettbewerben (sogenannten *Shared Tasks*) gehörten.

2.4.1.4. Rocchio/Nearest Centroid

Der *Rocchio-Klassifikator* (Rocchio (1971)) erhält seinen englisch ebenfalls geläufigen Namen *Nearest centroid classifier* nach dem Ansatz, die Dokumentvektoren jeweils einer Klasse des Trainingskorpus in einen Schwerpunktvektor zusammenzufassen. Die Bildung dieses Zentroiden erfolgt durch Addition aller Trainingsvektoren und anschließende Normalisierung auf die Anzahl der Trainingsbeispiele der Klasse und somit Konstruktion eines Durchschnittsvektors der Klasse in der Form

$$\vec{\mu}(c) = \frac{1}{\mid D_c \mid} \sum_{d \in D_c} \vec{v}(d) \tag{2.36}$$

Die Entscheidungsgrenze zwischen zwei Klassen verläuft dann jeweils exakt in der Mitte zwischen deren Schwerpunktvektoren, so dass ein zu klassifizierender Dokumentvektor $\vec{x}$ sich exakt auf dieser Grenze findet, wenn

$$|\vec{\mu}(c_1) - \vec{x}| = |\vec{\mu}(c_2) - \vec{x}| \tag{2.37}$$

Somit ist die als Entscheidungsgrenze fungierende Hyperebene definiert durch die auf ihr liegende Menge aller Punkte, für die gilt

$$\vec{w}^\top \vec{x} = b \tag{2.38}$$

wobei $\vec{w}$ der Normalenvektor zur auf dieser Ebene liegenden Punktmenge und b eine zu erlernende Konstante seien.

Die naheliegende Zuweisungsregel, einem Vektor eine Klasse c zuzuordnen, wenn

$$\vec{w}^\top \vec{x} > b \tag{2.39}$$

und $\bar{c}$, wenn

$$\vec{w}^\top \vec{x} \leq b, \tag{2.40}$$

entspricht einem linearen Klassifikator, bei dem

$$\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2) \tag{2.41}$$

sowie

$$b = 0,5 * (|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2) \tag{2.42}$$

Alternativ kann die Zuweisung zu der Klasse erfolgen, deren Schwerpunktvektor die größte Kosinusähnlichkeit zum Dokumentvektor aufweist.

2.4.2. Nichtlineare Klassifikatoren am Beispiel des K-nächste-Nachbarn-Klassifikators

Zwar weisen die bisher besprochenen Klassifikatoren zum Teil Transformationen auf wie die Logarithmierung der gewichteten Merkmale zur Erzeugung eines Polynoms (Naive Bayes) oder nichtlineare Erweiterungen wie die Kernelfunktionen der Support Vector Machine, sie haben jedoch die Gemeinsamkeit, dass sie als lineare Gleichungen formulierbar sind. Dieser Abschnitt stellt mit dem *K-nächste-Nachbarn-Klassifikator* einen Algorithmus vor, der nicht in dieser Weise formulierbar ist, sondern auf sequentiell durchgeführten Arbeitsschritten basiert.

Ähnlich wie der Rocchio-Klassifikator basiert der K-nächste-Nachbarn-Klassifikator (Cover and Hart (1967), Friedman et al. (2001)) auf einer geometrischen Interpretation der Trainingsvektoren einer Klasse als räumlich benachbart. Anstelle eines künstlich gebildeten Schwerpunktvektors werden hier jedoch sämtliche individuellen Dokumentvektoren eines Trainingskorpus mit dem zu klassifizierenden Vektor verglichen. Sind sämtliche Distanzen, etwa in Form der Kosinusähnlichkeit, ermittelt, entscheiden die namensgebenden k nächsten Nachbarn, also ähnlichsten Vektoren, per „Abstimmung“ über die Klassenzugehörigkeit des unbekannten Dokumentvektors. k ist hierbei ein zu trainierender Parameter. Da für den Klassifikationsvorgang stets sämtliche Trainingsvektoren mit dem Klassifikationsziel verglichen werden, existiert keine strenge Trennung zwischen Training und Klassifikation, die über die iterative Ermittlung von k hinausgeht. Die Notwendigkeit, sämtliche Trainingsbeispiele stets ad hoc mit dem unbekannten Vektor zu vergleichen, kann die Laufzeit dieses Klassifikationsverfahrens belasten. Diese wächst linear mit der Anzahl der Trainingsbeispiele, ist im Gegenzug jedoch unabhängig von der Anzahl der Zielklassen, die bei den bisher besprochenen Klassifikationsverfahren linearen Einfluss auf die Laufzeit hat.

Über die geschilderte Basisvariante einer Mehrheitsentscheidung hinaus besteht die Möglichkeit, einen auf den Kosinusähnlichkeiten der k nächsten Nachbarn basierenden Score für eine Klasse zu bilden, zu berechnen als

$$Score(c,d) = \sum_{d' \in S_k(d)} I_c(d') cos(\vec{v}(d'), \vec{v}(d)), \tag{2.43}$$

wobei $S_k(d)$ die Menge der k nächsten Nachbarn bezeichnet und $I_c(d') = 1$, wenn d Klasse c zugehörig ist, andernfalls $I_c(d') = 0$. Die Entscheidung fällt zugunsten der Klasse mit dem höchsten Score.

2.4.3. Multi-Klassen-Klassifikation in konventionellen Verfahren

Unter den bisher behandelten Klassifikatoren sind die Basisimplementationen der Support Vector Machine und Logistischen Regression inhärent nicht in der Lage, nichtbinäre Klassifikationsentscheidungen zu treffen: Über die Vorzeichen- beziehungsweise Sigmoidfunktion können einem vorgestellten Dokument lediglich zwei Werte, *Klasse c zugehörig oder nicht zugehörig*, alternativ *wahr oder falsch*, zugeordnet werden. Soll eine Zuordnung stattdessen in eine von n Kategorien erfolgen, werden in der Praxis n Klassifikatoren konstruiert und eingesetzt: Jeder dieser Klassifikatoren fungiert als sogenannter *One-vs-all-Klassifikator*, das heißt, er prüft auf die Zugehörigkeit des unbekannten Dokumentes zu einer bestimmten Klasse oder stattdessen zur nicht weiter aufgelösten Menge der verbleibenden Klassen. Nach der Durchführung sämtlicher Teilklassifikationen durch die einzelnen Klassifikatoren erhält das Dokument die Klassenzugehörigkeit mit dem höchsten Score, das heißt Abstand zur Hyperebene oder zum Schwellenwert

(so auch Pedregosa et al. (2011)). Der Naive-Bayes-Klassifikator hingegen liefert gegebenenfalls einen absteigend sortierbaren Score für jede Klasse, und die Klassifikatoren K-nächste-Nachbarn und Rocchio ermöglichen über die Kosinusähnlichkeit ebenfalls die Ermittlung eines klassenbezogenen Scores.

2.4.4. Künstliche neuronale Netze

Dieser Abschnitt beschreibt die Klassifikatorengruppe der sogenannten künstlichen neuronalen Netze. Die Grundlagen künstlicher neuronaler Netze, die als Klassifikatoren, Regressoren und Generatoren fungieren können, liegen in theoretischen Vorarbeiten von McCulloch und Pitts (*McCulloch-Pitts-Neuron*, McCulloch and Pitts (1943)) und Hebb (Hebb'sche Lernregel für neuronale Netze, Hebb (1949)) zur Modellierung biologischer Neuronen und deren Lernverhalten. Aufbauend auf diesen Vorarbeiten bildet das von Rosenblatt (Rosenblatt (1958)) definierte Modell des *Perzeptrons* die bis heute verwendete Grundlage künstlicher neuronaler Netze. Auf die Erläuterung der Funktionsweise des Perzeptrons und der grundlegenden Konzepte der auf ihm aufbauenden künstlichen neuronalen Netze folgen separate Abschnitte zu den in dieser Untersuchung verwendeten spezialisierteren Konzepte der *Embeddings* und *Convolutional Neural Networks*. Sofern nicht gesondert ausgewiesen, sind Terminologie und Definitionen stets Goodfellow et al. (2016) entnommen.

2.4.4.1. Das Perzeptron und die Architektur eines künstlichen neuronalen Netzes

Ein Perzeptron erhält in Anlehnung an das Modell des biologischen Vorbilds eine Menge von Eingabewerten x_i, die zu einer Menge von Ausgabewerten o_j, berechnet als

$$o_j = \sum_i w_{ij} x_i, \tag{2.44}$$

kombiniert werden, wobei w_{ij} einen spezifischen Gewichtungsvektor bezeichnet. Anstelle einer einfachen linearen Funktion wie in Gleichung 2.44 handelt es sich bei der Ausgabe eines Perzeptrons in aller Regel um eine nichtlineare Funktion ihrer gesamten Eingaben. Zu Beginn der Arbeiten an künstlichen Neuronen erfolgte dies in der Form

$$o_j = 1 \; wenn \; \sum_i w_{ij} x_i + b > 0, \tag{2.45}$$

wobei b den sogenannten Bias, mit negativem Vorzeichen einem Schwellenwert entsprechend, bezeichnet.

Des Weiteren kann die sigmoide Funktion

$$o_i = \frac{1}{1 + e^{-x_i}}, \tag{2.46}$$

die als logistische Funktion im nach ihr benannten Klassifikator in Unterabschnitt 2.4.1.2 vorgestellt wurde, zur nichtlinaren Modifikation oder als Schwellenwertfunktion für die Gesamteingabe der Einheit verwendet werden. Eine Funktion, die lineare Eingabewerte eines künstlichen Neurons in einen Ausgabewert in einem erwünschten Wertebereich transfomiert, wird *Aktivierungsfunktion* genannt. Im Laufe der Jahre wurden zahlreiche alternative Aktivierungsfunktionen entwickelt beziehungsweise für diesen Zweck entdeckt und anstelle der Sigmoidfunktionen verwendet, etwa der Tangens Hyperbolicus und die

in den letzten Jahren verbreitete *Rectified Linear Unit (ReLU)* (Hahnloser et al. (2000), Jarrett et al. (2009), Ramachandran et al. (2017)),

$$f(x) = max(0, x) \tag{2.47}$$

Eine Modifikation der ReLU-Funktion, die Werte ≤ 0 mit kleinen Koeffizienten gewichtet, ist unter der Bezeichnung *Leaky ReLU* bekannt (Maas et al. (2013)).

Die derart definierten künstlichen Neuronen, im Folgenden in Anlehnung an den englischen Sprachgebrauch (*units*) kurz *Einheiten* genannt, werden in sogenannten *Layern* oder *Schichten* angeordnet, die wiederum sequentiell angeordnet werden. Jede Einheit eines Layers kann Eingaben beliebig vieler Einheiten der vorhergehenden Schicht annehmen und Ausgaben an beliebige Einheiten der folgenden Schicht senden. Die erste dieser Schichten wird *Eingabeschicht* (*Input Layer*), die letzte *Ausgabeschicht* (*Output Layer*) genannt. Die zwischen diesen Schichten angeordneten Layer werden als *verdeckte Schichten* (*Hidden Layers*) bezeichnet. Ein solches Netzwerk wird *Multilayer Perceptron* genannt. Werden Aktivierungen im Netzwerk ausschließlich von der Eingabe- bis zur Ausgabeschicht transportiert, handelt es sich um ein *Feed Forward Network*. Netzwerke, die auch die Möglichkeit beinhalten, Aktivierungen in vorhergehende Schichten zu übertragen, werden als *rekurrente neuronale Netze* (*recurrent neural networks*), kurz *RNN*, bezeichnet. Letztere eignen sich besonders zur Verarbeitung von ihrer Natur nach sequentiellen Daten, wie sie etwa in Audiosignalen, Videos und natürlicher Sprache vorkommen. Rekurrente Netzwerke werden in dieser Untersuchung im Hinblick auf den Untersuchungsgegenstand, den Einfluss der Flexion auf Unigramm-Merkmale im Bag-of-Words-Modell, nicht verwendet und daher hier nicht weiter behandelt.

Die *Softmax-Funktion*,

$$\sigma(\alpha_j) = \frac{e^{\alpha_j}}{\sum_{k=1}^{K} e^{\alpha_k}} \tag{2.48}$$

wird regelmäßig zur Normalisierung der Werte α_j der Ausgabeeinheiten auf das gesamte Aktivierungspotenzial aus der letzten verdeckten Schicht verwendet, um eine proportionale Einordnung der Vorhersagestärke in einem nichtbinären Klassifikationsszenario zu ermöglichen. Diese wird in grober Näherung auch als Wahrscheinlichkeit einer Zuordnung aufgefasst.

Die Quantifizierung der Genauigkeit des neuronalen Modells erfolgt durch eine auf den Output-Layer angewandte *Verlust-* oder *Kostenfunktion* (englisch *cost function, loss function*). Eine für Multiklassenklassifikation übliche Verlustfunktion ist die *Categorical Cross-Entropy*,

$$CE = -\sum_{i}^{C} t_i log(f(s)_i), \tag{2.49}$$

wobei es sich bei $f(s)$ wiederum üblicherweise um die Softmax-Funktion (Gleichung 2.48) handelt.

Kostenfunktionen müssen stetig differenzierbar sein, um den eigentlichen Lernvorgang des neuronalen Netzes, die Berechnung der eingangs zufallsinitialisierten Gewichtungen der Übertragungen aus den einzelnen Einheiten durch das heute übliche *Backpropagation*-Verfahren (Rumelhart et al. (1986)) zu ermöglichen. Hierbei wird über die partiellen Ableitungen nach jeder Eingangsvariablen einer Einheit der Anteil dieser Einheit am entstandenen, durch die Kostenfunktion bemessenen, Fehler *zurückgeführt*. Diese

Rückführung erfolgt über alle Schichten des Netzes bis zur Eingabe mit einer zu wählenden *Lernrate*. Zu diesem Zweck ist es erforderlich, dass außer der Verlustfunktion auch sämtliche Aktivierungsfunktionen differenzierbar sein müssen. In der Praxis wird die ReLU-Funktion zu diesem Zweck durch

$$f(x) = ln(1 + e^x) \tag{2.50}$$

(bekannt als *Softplus*) approximiert.

2.4.4.2. Embeddings

Ein *word embedding* ordnet einer Wortform einen n-dimensionalen reellwertigen Vektor zu. Hierdurch kann eine Wortform, flektiert oder als Lemma, in einem Vektorraum verortet werden, in dem einander ähnliche Wortformen möglichst in räumlicher Nähe zu finden sind. Diese Ähnlichkeit kann hierbei als morphologisch, semantisch oder eine Kombination aus beidem interpretiert werden. Das Konzept, derartige Repräsentationen von Wortformen mit neuronalen Netzen zu erzeugen, wurde durch Bengio et al. (2003) entwickelt und mit dem Erfolg des frei verfügbaren *word2vec* (Mikolov et al. (2013)) verbreitet. Word2vec etablierte die Konzepte *CBOW* (*Continuous-Bag-Of-Words*) und *Skip-gram* als Trainingsziele bei der Erstellung solcher Repräsentationen: Das neuronale Netz wird im CBOW-Modell darauf trainiert, den möglichen Kontext eines bekannten Wortes vorherzusagen, und umgekehrt im Skip-gram-Modell auf die Benennung eines oder mehrere fehlender mittlerer Wörter bei bekanntem Kontext. Ein solches Optimierungskriterium erzwingt indirekt eine Erlernung von morphosyntaktischen Eigenschaften des Wortes und ihre Kondensation im repräsentierenden Vektor: Nur durch korrekte Berücksichtigung von Phrasenkongruenz in Genus, Numerus und Kasus etwa lassen sich

Wörter der offenen Wortklassen in einem deutschen Satz so ergänzen, dass dies als gelungen gewertet werden kann. Auch semantische Angemessenheit muss bei der Erfüllung beider Aufgaben berücksichtigt werden, was nur gelingen kann, wenn auch diese Informationen im zugrundeliegenden Vektor integriert wurden. *FastText* (Joulin et al. (2016a), Joulin et al. (2016b)) bezeichnet sowohl ein Embedding als auch ein darauf aufbauendes Klassifikationssystem, die beide ebenfalls frei verfügbar sind. Bei der Erstellung der Vektorrepräsentationen greift FastText im Training neben der eigentlichen Wortform auch auf deren wählbar aufgelöste N-Gramme zurück. Deren Berücksichtigung sollte in flektierenden Sprachen alle flektierten Formen und somit auch diejenigen, die nicht selbst im Trainingskorpus zu finden sind, zumindest indirekt repräsentieren. Die Architektur der verwendeten neuronalen Netze unterliegt einer stetigen Evolution mit schneller Generationenfolge, so dass das zu Beginn dieser Untersuchung weit verbreitete FastText in kurzer Zeit von RNN- (*ELMO*, Peters et al. (2018)) und später transformerbasierten Embeddings (*BERT*, Devlin et al. (2018)) abgelöst wurde.

2.4.4.3. Convolutional Neural Networks

Convolutional Neural Networks (LeCun et al. (1989)), kurz *CNN*, sind eine Erweiterung des Multilayer-Perceptrons zur Verarbeitung von Daten, die in Form n-dimensionaler Matrizen vorliegen oder in eine solche überführt werden können. Im Falle der ursprünglichen Hauptanwendung Bildverarbeitung sind dies etwa die Pixel eines Bildes, angeordnet in Spalten und Reihen. Ebenfalls prädestiniert für die Verarbeitung in einem CNN sind Textdaten in Form ihrer Embeddingvektoren: Hier können die Vektoren der laufenden Wortformen eines Textes in fixer Länge entweder in Spalten oder Reihen konkateniert werden, so dass die einzelnen Werte dieser Vektoren entsprechend in Reihen oder Spalten zu finden sind. Die Konvolutionsoperation

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \tag{2.51}$$

bewegt einen Kernel $K(m,n)$, der einer Art Sichtfenster entspricht, über eine zweidimensionale Matrix I. Jeder Wert der Matrix I wird mit dem Wert des über ihr befindlichen Kernelpunktes multipliziert und zum Wert der übrigen aktuell unter dem Kernel befindlichen Werte addiert und ergibt dann eine neue Stelle der Zielmatrix. Rückgabewert dieser Operationen über die gesamte Matrix I ist eine neue Merkmalsmatrix, englisch *feature map*, die aus kontextbasiert ausgewerteten und gewichteten Werten der Originalmatrix besteht. Als Convolutional Neural Network kann jedes Netzwerk bezeichnet werden, das mindestens einen Layer mit dieser Art Konvolutionsoperation enthält, wobei der Anzahl der Convolutional Layer prinzipiell keine Grenzen gesetzt sind. Convolutional Layer dienen der Zusammenfassung von Informationen mit ihrer Umgebung und erzeugen so neue, zusammengesetzte Merkmale. Am Beispiel der Bildverarbeitung können so etwa Kanten detektiert und Gesichtsbestandteile wie Augen erkannt werden, die zu komplexeren Einheiten wie Gesichtern abstrahiert werden. Ein denkbares Äquivalent in der Verarbeitung natürlicher Sprache ist die Zusammenfassung flektierter Formen zu ihrem Lemma. Die Kerneloperation ermöglicht nicht nur die Definition eines abstrakteren Merkmals, sondern auch dessen positionsunabhängige Verwendung. Üblicherweise werden die durch einen Convolutional Layer gebildeten Merkmale anschließend in einen regulären, vollständig verbundenen Layer überführt. Optional können noch sogenannte *Pooling-Layer* hinzugeschaltet werden, in denen die Ausgabe auf verschiedene Arten auf umgebungsbasierte, transformierte Werte *gepoolt* wird, um die Toleranz gegenüber moderaten Änderungen in der Eingabe- oder Vorgängerschicht zu erhöhen. Eine verbreitete Variante dieser Pooling-Operationen ist *Max-Pooling*, bei dem einem Neuron der höchste Wert seiner n Nachbarn zugeordnet wird.

Wie im vorangegangen Abschnitt erklärt, kondensieren Embeddings eine Reihe von morphosyntaktischen (etwa Numerus oder Tempus) wie semantischen (etwa das natürliche Geschlecht) Merkmalen einer Wortform in einen Vektor. Wird nun ein Eingabetext in die konkatenierten Embeddingvektoren jeder seiner laufenden Wortformen konvertiert, kann er als zweidimensionale m x n-Matrix aufgefasst werden, bei der m der Anzahl der Wörter entspricht, also jeder Vektor eine Zeile bildet, und n die Anzahl der Spalten, gleich der Dimension der Vektoren, in denen die Informationen zur Wortform gespeichert sind. Ein diese Matrix durchlaufender Kernel eines Convolutional Layers kann nun sowohl die für das Lernziel des neuronalen Netzes, etwa Klassifikation, relevanten Wörter (über die Reihen), als auch spezifisch die relevanten Aspekte jener Wörter (über die Spalten des Embeddingvektors) extrahieren beziehungsweise als Merkmale zusammenfassen.

3. Ressourcen

Das vorhergehende Kapitel beschrieb die grundsätzliche Architektur eines Klassifikationssystems und die zu seinem Training und Test benötigten Ressourcen sowie eine Reihe von Klassifikationsverfahren. Hieran anschließend werden in diesem Kapitel die Ressourcen benannt, die zu einer exemplarischen, korpusbasierten empirischen Untersuchung des Einflusses der Flexionsmorphologie in der Textklassifikation eingesetzt wurden. Es handelt sich hierbei um die Treebank *TübaDZ*, die zur Generierung eines Klassifikationsszenarios mit verschiedenen Korpusvariationen verwendet wurde, das freie Wörterbuch *Wiktionary* als ergänzende lexikalische Ressource, die Softwarebibliotheken *Scikit-learn*, *Keras* und *Tensorflow* sowie das Wortembedding *FastText*.

3.1. TübaDZ als Basis zur Erstellung eines Klassifikationsszenarios

Für die Weiterentwicklung und Evaluation von Klassifikationsverfahren steht grundsätzlich eine große Menge etablierter mono- und multilingualer Korpora zur Verfügung. Unter diesen gilt das multilinguale, auch deutschsprachige Nachrichtentexte enthaltende Reuters Corpus Volume 1 (Russell-Rose et al. (2002)) mit rund 810.000 Nachrichtentexten als verbreiteter und fortlaufend aktualisierter (etwa Schwenk and Li (2018)) Benchmark

für die Klassifikation standardsprachlicher Texte. Daneben wurden über die Jahre der Entwicklung der Textklassifikation zahlreiche Korpora ad hoc für spezifische Untersuchungsaspekte editiert (Leopold and Kindermann (2002)). Derartige Korpora mussten jedoch ebenso wie das Reuters RCV1 in Ermangelung einer Goldstandardannotation mit morphosyntaktischen Informationen als Basis für die vorliegende flexionsmorphologiezentrierte Untersuchung sämtlich verworfen werden. Die alternativ denkbare Erzeugung dieser Annotationsschicht auf Standardkorpora mithilfe von Annotationswerkzeugen wie Part-of-Speech-Taggern und Lemmatisierern (etwa dem TreeTagger, Schmid (1995), Schmid (1999)) war aus methodischer Sicht zu verwerfen: Die Evaluation der Reaktion von Klassifikatoren auf flexionsmorphologische Phänomene, die ihrerseits im Korpus von Software detektiert und annotiert wurden, unterliegt den schwankenden, korpusspezifischen Fehlerquoten solcher Werkzeuge. Deren Ausmaß und Verteilung ist bereits in der Literatur umstritten (beispielhaft Giesbrecht and Evert (2009)) und qualitativ und quantitativ mangels Referenzrahmen nicht szenariospezifisch bestimmbar. Im Zuge der Verwendung in Merkmalsauswahl- und Klassifikationsverfahren über Text- und Klassengrenzen hinweg ist die Weiterentwicklung dieser Residuen nicht nachzuverfolgen und verwässert als Rauschen gewonnene Erkenntnisse. Bei den Treebanks zur deutschen Sprache *NEGRA* (Skut et al. (1997)), *TIGER* (Brants et al. (2002)) und *TübaDZ* (Telljohann et al. (2004)) handelt es sich dagegen um etablierte Korpora, deren Entwicklung primär auf die Bereitstellung von Ressourcen für syntaxtheoretische Untersuchungen ausgerichtet war. Treebanks erschienen durch ihre morphosyntaktische Annotationsschicht mit Part-of-Speech-Tags, Lemmata und morphologischen Informationen als denkbare Alternative zur Verwendung der primär für Klassifikationsforschung verwendeten unannotierten Standardkorpora. Die dritte Alternative einer eigenen, manuellen morphosyntaktischen Annotation eines bestehenden Standardklassifikationskorpus in hinreichendem Umfang erschien angesichts des zu erwartenden hohen Aufwandes für die qualitätsgesicherte Annotation von Part-of-Speech-Tags, Morphologie und Lemmata in hinreichender Menge als illusorisch: Brants et al. (2002) beziffern die Annotations-

geschwindigkeit auf dieser Ebene auf NEGRA mit etwa 1.300 Tokens pro Stunde pro Annotator. Die Begründung des Verzichts auf etablierte Korpora bei der Durchführung von Experimenten in einer etablierten computerlinguistischen Disziplin wie der Textklassifikation ist im Fall dieser Untersuchung somit primär in der Unabdingbarkeit morphosyntaktischer Goldstandardannotationen des Korpus zu sehen. Da es sich bei Treebanks nicht (mehr) um Textsammlungen in Form von Einzeldokumenten handelt, diese jedoch die Grundlage ihrer Zusammenstellung bildeten, war die Möglichkeit zu untersuchen, durch erneute Aufteilung einer annotierten Treebank in ihre Einzeltexte an ein annotiertes Nachrichtentextkorpus zu gelangen. Dieses wäre sodann in inhaltliche Klassen unterteilt in ein Klassifikationsszenario zu konvertieren. Der Aufwand für die Annotation der Texte mit einer Klassenzugehörigkeit bei vorhandener morphosyntaktischer Annotation beträgt einen Bruchteil des Zeitbedarfs für die umgekehrte Herangesehensweise, die morphosyntaktische Annotation einzelner Texte eines etablierten, aber unannotierten, Klassifikationskorpus.

Grundlage aller drei Treebanks sind serialisierte deutschsprachige Nachrichtentexte aus den 1990er-Jahren. NEGRA als älteste der drei Treebanks (Weiterentwicklung 2006 eingestellt) wurde im Vergleich zu TIGER und TübaDZ als mit 355.100 Tokens, davon lediglich 60.000 goldstandardannotiert quantitativ nicht konkurrenzfähig ausgeschlossen. TIGER weist mit rund 890.000 Tokens laufendem Text nur gut die Hälfte des Umfangs der aktuellsten und bis in jüngere Zeit (finale Edition 2018) entwickelten Treebank TübaDZ auf. Zusätzlich zum größeren Umfang bot TübaDZ zum Beginn der Untersuchung 2016 gegenüber TIGER den erheblichen technischen Vorteil einer automatisierbaren Trennbarkeit in die ursprünglichen Nachrichtentexte: Die für den Aufbau eines Klassifikationskorpus benötigte Zuordnung der Sätze zu einzelnen Texten ist im ursprünglichen TIGER nicht notiert und nur teilautomatisch mit manueller Nachbearbeitung rekonstruierbar. Eine entsprechende Ergänzung zu TIGER erfolgte erst nach Beginn dieser Untersuchung. Dieser Umstand ergänzte die Überlegung, dass die doppelte Menge laufenden Textes in

TübaDZ in Hinblick auf Zipfs Gesetz zwar keine exakt proportional bessere Klassifikationsleistung induzieren, wohl aber eine größere Anzahl Einzeltexte verfügbar machen und somit größere Gestaltungsfreiheit im Klassifikationsszenario ermöglichen würde. Somit fiel die Wahl zur Grundlage der Erstellung eines Klassifikationsszenarios auf die exakt 1.883.396 Tokens (Edition 2016) umfassende Treebank TübaDZ, deren fortlaufend im CONLL-Format (Buchholz and Marsi (2006)) zeilenweise annotierten Tokens in 3.643 Texte konvertiert werden konnten. Diese Texte dienten als Ausgangsbasis eines neu erstellten Klassifikationsszenarios. Für die manuelle Annotation einer äquivalenten Menge Material wäre nach den Zahlen von Brants ein Annotationsaufwand von gut 1.500 Stunden pro Annotator zu veranschlagen gewesen. Diese entsprechen bei der zur Qualitätssicherung erforderlichen minimal zweifachen Annotation etwa 370 Arbeitstagen, die im Rahmen dieser Untersuchung nicht geleistet werden konnten. Das CONLL-Format beinhaltet die folgenden Informationen in den ersten sechs Spalten von links nach rechts gelesen: Token-ID, Wortform, Lemma, POS-Tag (Hauptkategorien), STTS-POS-Tag sowie Informationen zur Morphologie. Das für Deutsch etablierte Tagset STTS (Schiller et al. (1995), Thielen (1999)) wurde als Ausgangspunkt für die erhobenen Korpusstatistiken und die vorgenommenen Korpusmodifikationen verwendet.

Eine erste manuelle Sichtung von 100 zufällig ausgewählten Texten aus TübaDZ etablierte als Startpunkt für das Klassifikationsszenario eine initiale Menge von acht Inhaltskategorien „Konflikte Ausland“, „Kriminalität“, „Kultur“, „Panorama“, „Politik“, „Sport“, „Umwelt“ und „Wirtschaft“. Die 3.643 Nachrichtentexte wurden sodann zu gleichen Teilen von zwei Annotatoren diesen Kategorien zugewiesen. Das Inter-Annotator-Agreement, gemessen in *Cohen's Kappa* (Cohen (1960)), wurde auf je 100 von beiden Annotatoren doppelt annotierten, zufällig ausgewählten Texten erhoben und betrug $0,58$.

Tabelle 3.1 (Seite 42) zeigt die Verteilung der Texte auf die einzelnen Kategorien.

Kategorie	Texte
Konflikte Ausland	351
Kriminalität	309
Kultur	778
Panorama	671
Politik	818
Sport	210
Umwelt	166
Wirtschaft	340

Tabelle 3.1.: Anzahl der den Kategorien zugeordneten Texte

Auch unter Berücksichtigung der begründeten Erkenntnis, dass durch den Bedarf an einer morphosyntaktischen Goldannotationsschicht in jedem Falle eine Treebank als Basis für die Neuerstellung eines Klassifikationsszenarios gewählt werden musste, erscheint an dieser Stelle die Frage nach der Begründung von Anzahl, Benennung und Besetzung der Kategorien des so gebildeten Korpus legitim. Im Hinblick auf den Untersuchungsgegenstand und damit den Verwendungszweck des neugeschaffenen Korpus soll die folgende Argumentation die Zurückstellung dieser als semantikbezogen aufgefassten Aspekte begründen: Gegenstand der Untersuchung ist der Einfluss der Flexion der deutschen Sprache auf den Klassifikationsvorgang, das heißt, der Fähigkeit verschiedener Algorithmen, Dokumente eines vorliegenden Korpus vorab bestimmten Kategorien zuzuweisen. Der absolut erreichbare Klassifikationserfolg ist dabei abhängig von der Ausdrucksstärke des jeweiligen Algorithmus und seiner Kompatibilität mit dem Korpusszenario, der Anzahl und Trennschärfe der Kategorien sowie der Anzahl der Trainings- und Testbeispiele und ihrer Distributionen. Diese absolute Klassifikationsleistung als Funktion der aufgezählten Variablen ist jedoch gerade nicht Gegenstand der Untersuchung. Die Veränderungen der im Klassifikationsprozess bedeutungstragenden Wortformen durch Flexionsprozesse führen zu Verschiebungen der Verteilung der Wortformen auf Korpusebene, die von Klassifikationsalgorithmen interpretiert werden. Ohne einschlägige manuelle Eingriffe durch Annotatoren besteht grundsätzlich kein Anlass, von einer quantitativ relevanten Verschie-

bung dieser Verteilungen auf der neu eingeführten Ebene der Kategorie auszugehen. Die Abbildungen 3.1 und 3.2 zur Rangfolge der jeweils 50 häufigsten Types zeigen einen mit dieser Annahme übereinstimmenden hohen Kongruenzgrad sowohl der acht Kategorien untereinander als auch im Vergleich zur Gesamtverteilung auf Korpusebene.

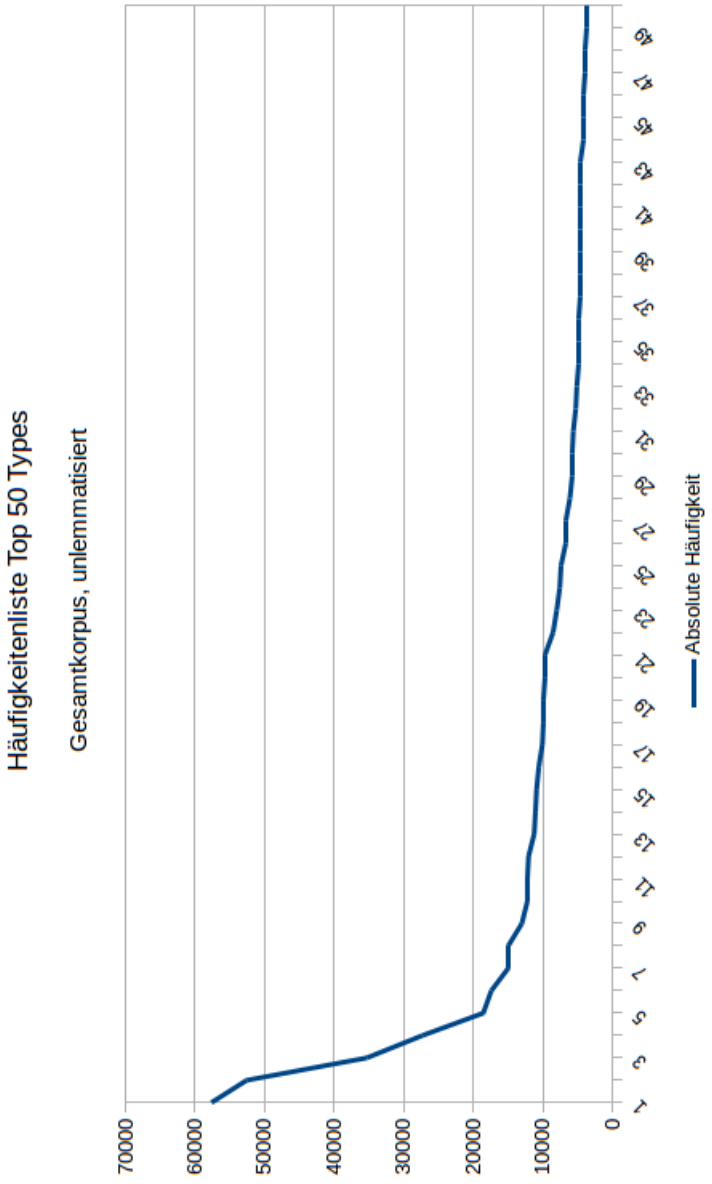


Abbildung 3.1.: Zipfkurve der 50 häufigsten Types in TübaDZ, gesamtes Korpus

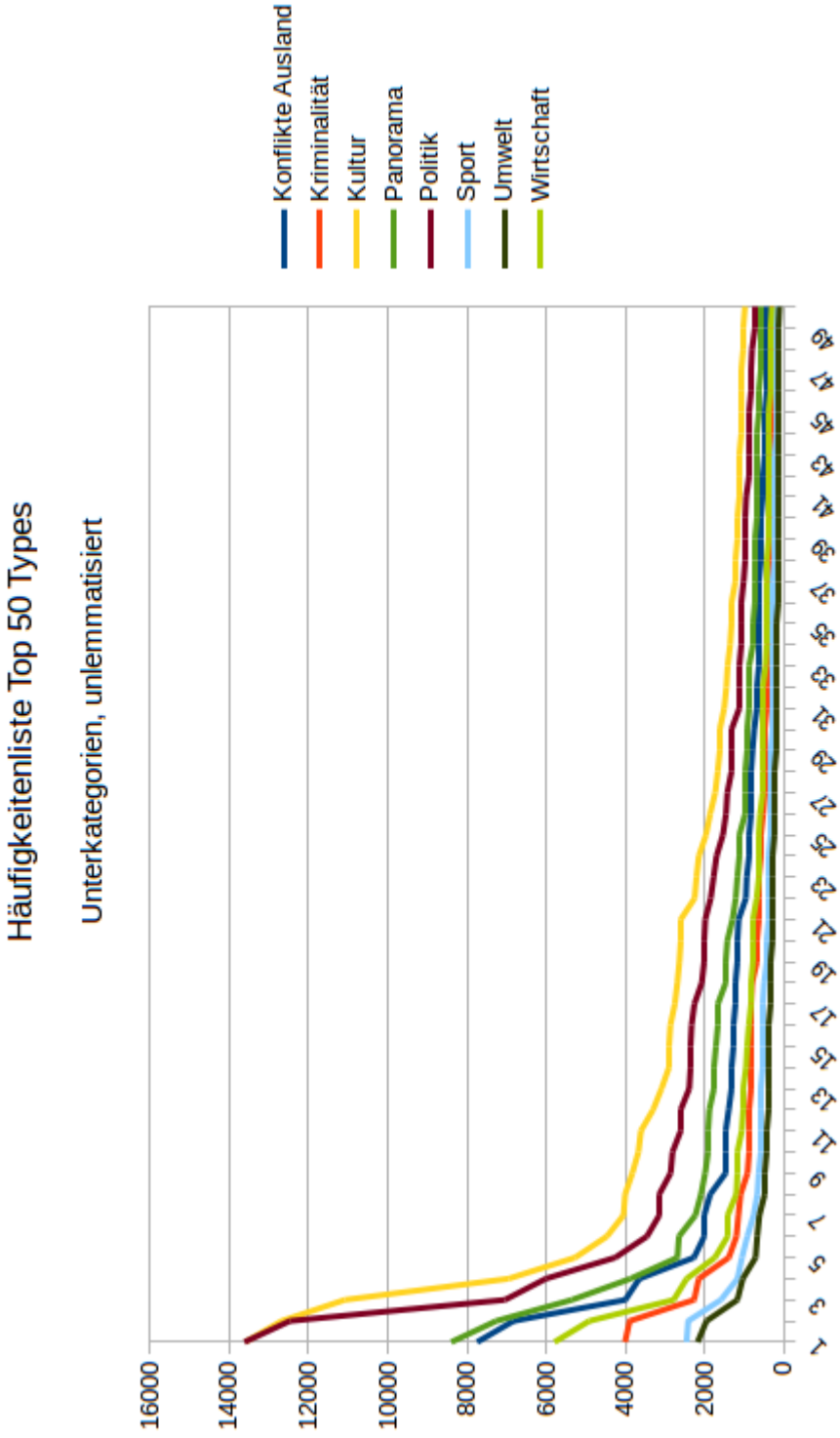


Abbildung 3.2.: Zipfkurve der 50 häufigsten Types in TübaDZ, alle Kategorien

Weitere kategorieübergreifende starke Übereinstimmungen zeigen sich in den in Unterkapitel 5.1 erhobenen Verteilungen der offenen Wortklassen. Ebenfalls einen hohen Kongruenzgrad zeigen die Entwicklung der kategoriespezifischen Merkmalsräume in Hinblick auf die Quote der Wortformen pro Lexem und die semantisch-morphologische interne Dichte nach Kosinusabstand (siehe ebenfalls Kapitel 5.1). Die gute Vergleichbarkeit der Kategorien unter diesen vier Aspekten begründet somit die fortan geltende Annahme, sowohl nach Vergleich der Kategorien untereinander als auch mit der Gesamtkorpusebene keine dem Untersuchungszweck zuwiderlaufenden statistischen Verzerrungen durch die Anzahl und Aufteilung der Kategorien zu erwarten. Darüber hinaus soll auch das Erreichen den Erwartungswert deutlich überschreitender Ergebnisse in den Klassifikationsexperimenten des Unterkapitels 5.2 als weiteres Indiz für eine gewisse statistische Objektivierbarkeit im Sinne einer Reproduzierbarkeit der Kategorienzuteilung gewertet werden.

3.2. Wiktionary als Wörterbuch

Das freie Online-Wörterbuch der Wikimedia Foundation, das 2002 gegründete *Wiktionary* (http://de.wiktionary.org), wurde zur Erstellung einer zusätzlichen lexikalischen Ressource zu den offenen Wortklassen genutzt. Die durch das freie Editionsprinzip über die lange Betriebsdauer abgesicherte lexikalische Qualität machte Wiktionary sowohl zum Untersuchungsziel an sich (etwa Meyer and Gurevych (2010), Meyer and Gurevych (2012b)) als auch zur Ressource für Projekte diverser computerlinguistischer Disziplinen. Letztere weisen eine große Bandbreite, von Phonetik (Schlippe et al. (2010)) über Synonymdetektion (Weale et al. (2009)) bis zu Ontologien und lexikalischer Semantik auf (Zesch et al. (2008), Meyer and Gurevych (2012a)). Bemühungen, Wiktionary speziell zur Extraktion morphologischer Informationen des Deutschen zu nutzen, finden sich bei Sennrich and Kunz (2014)) sowie Kirov et al. (2016).

Da Wiktionary somit als wissenschaftlich gesicherte, umfangreiche und frei verfügbare (Lizenz CC-by-SA, bezieht sich auf die Artikel als solche) Ressource gelten kann und da die Wörterbucheinträge einen hohen Grad an extraktionsfreundlicher Standardisierung aufweisen, lag es nahe, eine neue, gegenüber etwa ZMorge aktualisierte Extraktion vorzunehmen. Hierzu wurden im Oktober 2017 die Flexionstabellen sämtlicher verfügbarer Substantive, Adjektive und Verben extrahiert und in einem zeilenbasierten CSV-Format gespeichert. Dem jeweiligen Lemma folgen in diesem Format die deklinierten beziehungsweise konjugierten Formen spaltenweise an fester Position. Diese Auflistung nimmt einen gewissen Grad an Redundanzen durch Synkretismen in Kauf, da durch fixe Spaltenzahl und -position eine automatische Lesbarkeit gewährleistet werden kann. Das so gewonnene automatisch lesbare Wörterbuch enthält 61.181 Substantiv-, 9.058 Adjektiv- und 6.417 Verblemmata. Nach Abzug der Synkretismen finden sich 155.403 deklinierte Formen der Substantive, 17.658 deklinierte Adjektivformen und 34.554 konjugierte Verbformen im Wörterbuch. Hieraus ergibt sich ein Gesamtumfang von 207.615 Wortformen in den drei Teillisten. Einige Wortformen erscheinen als Homographen in mehr als einer dieser Listen. Zur Thematik der Homographien, ihrer Handhabung in der Klassifikation und der Motivation, Wiktionary als externe Ressource besonders bei der Untersuchung der Homographie zu berücksichtigen, siehe Unterkapitel 4.5 und 5.1.

3.3. Klassifikationssoftware

Sämtliche Klassifikationsexperimente wurden mit Hilfe der Softwarebibliotheken *Scikit-learn* und *Keras* duchgeführt, die in den folgenden Abschnitten kurz beschrieben werden. Unterstützende Skripte zur Vorbereitung und Konvertierung der Korpora und Wörterbücher, der Parametrisierung und automatischen Evaluation der Experimente sowie der Konvertierung der Testergebnisse in darstellbare Formate wurden in der Programmiersprache *Python 2.7* unter dem Betriebssystem Ubuntu 16.04 erstellt; diese Aufgaben

wurden nach Bedarf ergänzend auch von Skripten in der Shell-Sprache *Bash* übernommen.

3.3.1. Scikit-learn

Die auf der Programmiersprache Python basierende Softwarebibliothek *Scikit-learn* (Pedregosa et al. (2011)) stellt eine Vielzahl von Softwarepaketen unter anderem für die Disziplinen des maschinellen Lernens Clustering, Klassifikation und Regression bereit. Sie setzt dabei auf die Python-Bibliotheken *SciPy* (http://www.scipy.org, Virtanen et al. (2020)) und *NumPy* (http://numpy.org, Oliphant (2006)) auf und ist unter einer Vielzahl von Umgebungen lauffähig. Die bereitgestellten Implementierungen entsprechen gängigen Interpretationen der zugrundeliegenden Algorithmen und bieten diverse Konfigurationsmöglichkeiten.

3.3.2. Keras

Keras (Chollet et al. (2018)) ist eine ebenfalls auf Python basierende Softwarebibliothek, die eine leicht bedienbare API zu einer darunterliegenden, hardwarenäher implementierten Bibliothek für Deep Learning zur Verfügung stellt. Obwohl dieser Unterbau prinzipiell frei gewählt werden kann, hat sich in den letzten Jahren eine Kombination von Keras mit *TensorFlow* (Abadi et al. (2016)) etabliert, die etwa über die Repositorien von Ubuntu standardmäßig zur Verfügung gestellt wird und die auch in dieser Untersuchung zum Einsatz kam. Keras erlaubt eine baukastenartige, schnelle Erstellung selbst komplexer neuronaler Architekturen aus den grundlegenden Vorverarbeitungsschritten, Layern, Verlustfunktionen und Evaluationen. Ein Großteil der in Keras verfügbaren Funktionen wird für den Entwickler abstrahiert über TensorFlow gegebenenfalls unter Nutzung von

Spezialhardware ausgeführt. Auch Keras setzt auf eine starke Interaktion mit SciPy und NumPy.

3.3.3. FastText

Das in Unterabschnit 2.4.4.2 vorgestellte Embedding FastText wird von Facebook AI unter http://fasttext.cc in derzeit (März 2021) 147 Sprachen zur Verfügung gestellt. Bei den verfügbaren vortrainierten Modellen handelt es sich um 300-dimensionale, auf das CBOW-Kriterium optimierte Vektoren, die auf den Korpora von Common Crawl (http://commoncrawl.org) und den Einträgen der Wikipedia der jeweiligen Sprache trainiert wurden. Zum Zeitpunkt der Beschaffung im September 2018 enthielt das deutschsprachige Embedding 1.477.289 absteigend nach Häufigkeit sortierte Types nebst ihren Vektoren. Diese Fast-Text-Vektoren wurden als Eingabe für die Merkmalsraumuntersuchungen (siehe Abschnitt 5.1.3) sowie die in Unterkapitel 5.3 beschriebenen Klassifikationsexperimente verwendet.

4. Flexionsmorphologie in der Textklassifikation

Wie in Kapitel 2 dargestellt, bilden die Basis für Training und Einsatz eines Textklassifikators unabhängig von Vorverarbeitung, Merkmalsauswahlverfahren und Klassifikationsmethode stets die einen Text konstituierenden Wortformen. In Sprachen wie der deutschen ändern Wortformen zur Kodierung von Informationen ihre Form. Der Vorgang, dass sich Wortformen aus syntaktischen oder semantischen Erfordernissen verändern, wird Flexion genannt, die Lehre von den inneren und äußeren Systematiken dieser Veränderungen Flexionsmorphologie. Die vorliegende Arbeit beschränkt sich bei der Untersuchung, die Flexion auf die Textklassifikation nach Themen ausübt, auf Unigramme, also einzelne Wortformen. Der Umstand, dass ein Lexem durch Flexionsvorgänge in verschiedenen Formen erscheint, also durch verschiedene Zeichenketten repräsentiert wird, ist für ein unigrammbasiertes Klassifikationsmodell sowohl in Training als auch Evaluation und Klassifikation im produktiven Einsatz problematisch, da der Zusammenhang zwischen flektierter Form und zugrundeliegendem Lexem nicht ohne Weiteres ersichtlich ist: Die statistischen Zusammenhänge zwischen Lexem und Klasse werden durch Flexion über- oder unterschätzt. Ziel dieses Kapitels ist unter diesem Gesichtspunkt die Darstellung relevanter Aspekte der Flexionsmorphologie der offenen Wortklassen nebst einer empirischen Analyse auf dem Korpus zur Vorbereitung der experimentellen Untersuchungen in Kapitel 5.

Diese Studie strukturiert die Behandlung der flexionsmorphologischen Phänomene der Unterteilung im „Grundriss der deutschen Grammatik: Das Wort“ (Eisenberg (2020)) folgend. Alternative grundlegende Darstellungen der deutschen Grammatik finden sich in „Duden: Die Grammatik“ (Wöllstein (2016)) und im „Grammatischen Informationssystem“ des IDS Mannheim (Zifonun et al. (1997), abrufbar unter http://grammis.ids-mannheim.de). Eine vergleichende Diskussion dieser Systematiken konnte angesichts des Fokus dieser Arbeit zurückgestellt werden: Relevant für einen maschinellen Klassifikator ist weniger die innere Systematik der offenen Wortklassen als ihre empirischen Erscheinungen in Form von Zeichenketten in maschinell lesbarem Text.

Eisenberg übernimmt die traditionelle Aufteilung der Flexionsmorphologie in Deklination (Substantive, Adjektive, Pronomina und Artikel) und Konjugation (Verben). Zur Flexionsmorphologie zählt Eisenberg des Weiteren die Behandlung der trennbaren Verbpartikeln, da deren Verhalten von Konjugationsvorgängen abhängt. Dieser Unterteilung, ergänzt um das Phänomen der flexionsbedingten Homographie, folgt auch diese Studie.

Die Phänomene Deklination und Konjugation werden jeweils für sich behandelt, unterteilt in die Unterkapitel 4.1 Substantivdeklination, 4.2 Adjektivdeklination, 4.3 Konjugation und 4.4 trennbare Verbpartikeln. Im Hinblick auf Deklination geht diese Studie davon aus, dass bei einer Klassifikation nach Themen oder Inhalten nur die Wörter der offenen Wortklassen eine Rolle spielen, so dass die flektierenden Wortarten Pronomina und Artikel nicht betrachtet werden. Das Phänomen der flexionsbedingten wortklassenübergreifenden Homographie wird in Unterkapitel 4.5 behandelt. In jedem der genannten Unterkapitel wird zuerst ein kurzer Überblick über die Flexionsphänomene gegeben, gefolgt von einer quantitativen Erhebung auf dem Korpus zur Bestimmung empirischer Relevanz. In Unterkapitel 4.6 schließen sich Überlegungen zu den Auswirkungen der Flexionsvorgänge auf die Merkmalsauswahl und einzelne Klassifikatoren als Grundlage der experimentellen Untersuchungen von Kapitel 5 an.

4.1. Substantivdeklination

Das Flexionsparadigma des deutschen Substantivs enthält in der gegenwärtigen Sprachstufe acht Positionen für vier Kasus jeweils in Singular und Plural (Eisenberg (2020:167)). Die Formen an den acht Positionen in diesem Paradigma werden je nach Deklinationsklasse unterschiedlich stark ausdifferenziert; insbesondere beim Kasus gibt es zahlreiche Synkretismen. Eine Ursache hierfür ist, dass das Deutsche, wie auch andere germanische Sprachen, die in ihrer historischen Entwicklung hierbei schon weiter fortgeschritten sind (etwa Englisch und Dänisch), zum Abbau von Kasusmarkierungen tendiert, insbesondere bei Dativ, Akkusativ und auch Genitiv (ibd.). Eisenberg benennt vier Deklinationstypen mit je zwei Unterklassen. Grundsätzlich stehen zur Markierung von Kasus und Numerus die Endungen *e*, *n*, s und *r* zur Verfügung und werden in den einzelnen Klassen spezifisch eingesetzt. Mitunter existieren zwei unterschiedliche Stämme für Singular und Plural, wobei der Pluralstamm aus dem durch Umlautung modifizierten Stamm des Singular erzeugt wird. An den jeweiligen Stamm treten dann in der Regel die genannten Endungen; nur wenige Substantive bilden einen endungslosen Plural, der formgleich mit dem Nominativ Singular ist. Hierzu gehören etwa die Substantive *Eimer, Esel, Leiden* und *Wagen* (Beispiele aus Eisenberg(2020:172)). Tabelle 4.1 zeigt den unterschiedlichen Grad an Synkretismus in den Deklinationsklassen durch Aufzählung der unterschiedlichen Vorkommen von Suffixen im Paradigma[1]:

Aus den vorstehend aufgezählten Morphemen können kombiniert die folgenden Suffixe zur Markierung von Kasus und Numerus zusammengesetzt an den gegebenenfalls umgelauteten Stamm gesetzt werden: *e, es, s, en, er, ern* und *n*. Aus der vorstehenden Tabelle ergibt sich, dass ein Substantiv je nach Klasse im einfachsten Fall (Klasse 4,

[1] An dieser Stelle kann eine weitergehende Diskussion der inneren Systematik der Substantivdeklination, etwa der Frage, inwieweit es sich hier um agglutinierende oder fusionierende Mechanismen handelt, unterbleiben: Die Fragestellung der Arbeit bedingt eine Fokussierung auf die letztlich erzeugten Formen aus der Perspektive eines Klassifikators unabhängig vom Grund ihres Zustandekommens.

Klasse	Gruppe	Suffixe A	Verteilung	Suffixe B	Verteilung
1	M und N, s*	-,(e),(e)s,en	2+4+1+1	-,(e),(e)s,er,ern	2+3+1+1+1
2	M, schwach	en,(en),-	5+2+1	n,(n),-	5+2+1
3	M und N, g**	en,-,(e)s,(e)	4+2+1+1	n,-,s	4+3+1
4	F	en,-	4+4	-,e,en	4+3+1

Tabelle 4.1.: Flexionsklassen mit Synkretismen, *stark, **gemischt

Feminina, Gruppe a) zwei und im komplexesten Fall (1b, Maskulina und Neutra, stark) fünf Formen bildet.

	Singular	Plural
Nominativ	Burg	Burgen
Genitiv	Burg	Burgen
Dativ	Burg	Burgen
Akkusativ	Burg	Burgen

Tabelle 4.2.: Beispiel Substantivdeklination Feminina, keine Umlautung

	Singular	Plural
Nominativ	Kind	Kinder
Genitiv	Kind(es)	Kinder
Dativ	Kind(e)	Kindern
Akkusativ	Kind	Kinder

Tabelle 4.3.: Beispiel Substantivdeklination Maskulina und Neutra, stark

Somit besteht die Möglichkeit, dass ein Substantiv aus syntaktischen und semantischen Gründen in zwei bis fünf unterschiedlichen Formen, also Types, im Korpus auftritt. Der folgende Abschnitt stellt die empirische Verteilung der Substantive und ihrer Kategorien in TübaDZ dar.

4.1.1. Korpusanalyse TübaDZ

341.936 Tokens, entsprechend rund 21,1% der Tokens im Korpus, sind mit „NN“ und somit als Substantive getaggt (die folgenden Betrachtungen berücksichtigen als „NE“ getaggte Tokens nicht). Bei diesen 341.936 Tokens handelt es sich um Mehrfachvorkommen von 78.997 Types, bei denen es sich um flektierte Formen von 65.525 Lexemen handelt. Diese 65.525 Lexeme stellen mit 54,88% eine absolute Mehrheit der im Korpus vorkommenden 113.922 Lexeme und treten im Schnitt in 1,21 flektierten Formen auf. Dieser Abschnitt untersucht nach der somit offensichtlichen quantitativen Relevanz der Wortklasse Substantive im Korpus die empirische Verteilung der vorstehend benannten Flexionsphänomene.

4.1.1.1. Kasus

Tabelle 4.4 zeigt die Häufigkeiten sämtlicher möglicher Kombinationen des Auftretens eines Lexems in einem, zwei, drei oder allen vier Kasus absteigend geordnet. Neben den einzelnen Kasus treten sämtliche möglichen zwei- und dreistelligen Kombinationen (sechs und vier Möglichkeiten) sowie alle vier Kasus in Kombination zumindest gelegentlich auf, große Unterschiede in der Verteilung sind offensichtlich:

70,5% der Lexeme kommen nur in einem einzigen Kasus vor, 15,5% in zwei Kasus, 7,71% in drei und lediglich 6,3% in allen Kasus. Gleichzeitig sind die Kasus Nominativ, Dativ und Akkusativ in zumindest ähnlicher Größenordnung jeweils exklusiv vertreten.

Tabelle 4.5 zeigt die Verteilung der Kasus aus der Perspektive der Types:

75,9% der Types fungieren im Korpus in nur einem Kasus, 12,57% für zwei, 7,2% für drei, und nur 3,5% für alle vier. Zusammengefasst fungieren lediglich 23,7% der Types für mehr als einen und 10,7% für drei oder vier Kasus. Auch aus Perspektive der Types

Kasus	Auftreten	Entspricht
Nominativ	15.475	23,6%
Dativ	13.836	21,1%
Akkusativ	11.559	17,6%
Genitiv	5.392	8,2%
Nominativ, Akkusativ, Dativ, Genitiv	4.154	6,3%
Nominativ, Akkusativ, Dativ	2.889	4,4%
Nominativ, Dativ	2.589	4,0%
Nominativ, Akkusativ	2.467	3,8%
Akkusativ, Dativ	2.205	3,4%
Nominativ, Genitiv	1.088	1,7%
Dativ, Genitiv	930	1,4%
Nominativ, Dativ, Genitiv	910	1,4%
Akkusativ, Genitiv	755	1,2%
Nominativ, Genitiv, Dativ	679	1,0%
Akkusativ, Dativ, Genitiv	596	0,91%
Nominativ, Akkusativ, Dativ, Genitiv, *	1	0,002%
Summe	65.525	100%

Tabelle 4.4.: Auftreten Kasus Substantive, Lexeme

zeigt sich somit ein ähnliches Muster der Kasusverteilungen wie auf Lexemebene. Zusammengefasst lassen der geringe Anteil der in mehreren oder gar allen Kasus vertretenen Lexeme und die geringe Ausschöpfung des Reduktionspotenzials durch Synkretismen das Potenzial für Lernschwierigkeiten und Auswirkungen auf den Klassifikationsvorgang durch Kasusdeklination erheblich erscheinen.

4.1.1.2. Numerus

Die Auszählung der Verteilung des Numerus bei den Lexemen (Tabelle 4.6) zeigt einen Überhang von Singular zu Plural im Verhältnis von fast 3:1. Der Umstand, dass ähnlich zum Muster bei den Kasus lediglich 12,6% der Lexeme in beiden Numeri vorkommen, diese einzeln jedoch beide in relevantem Anteil, lässt analog ein Potenzial für statistische Verzerrungen vermuten.

Kasus	Auftreten	Entspricht
Nominativ	18.749	23,7%
Dativ	18.036	22,8%
Akkusativ	13.872	17,6%
Genitiv	9.289	11,8%
Nominativ, Akkusativ, Dativ	3.833	4,9%
Nominativ, Akkusativ	3.300	4,2%
Nomivativ, Akkusativ, Dativ, Genitiv	2.782	3,5%
Nominativ, Dativ	2.628	3,3%
Akkusativ, Dativ	2.392	3,0%
Nominativ, Genitiv	930	1,2%
Nominativ, Akkusativ, Genitiv	908	1,1%
Dativ, Genitiv	689	0,87%
Akkusativ, Genitiv	641	0,81%
Nominativ, Dativ, Genitiv	525	0,67%
Akkusativ, Dativ, Genitiv	422	0,53%
Nominativ, Akkusativ, Dativ, Genitiv und *	1	0,001%
Summe	78.997	100%

Tabelle 4.5.: Auftreten Kasus Substantive, Types

Numerus	Auftreten	Entspricht
Singular	42.333	64,6%
Plural	14.851	22,7%
Singular und Plural	8.266	12,6%
Singular, Plural und *	34	0,05%
Nur *	24	0,04%
Plural und *	12	0,02%
Singular und *	5	0,008%
Summe	65.525	100%

Tabelle 4.6.: Auftreten Numerus Substantive, Lexeme

Mit Tabelle 4.7 wird erneut die Perspektive gewechselt und das Auftreten der 78.997 Types in den beiden Numeri dargestellt. Klar erkennbar ist hier nun der geringe Anteil an Numerus-Synkretismen, was der geringen Zahl endungsloser Plurale (siehe Einleitung) im Wortschatz geschuldet ist, so dass sich ein im Vergleich zu den Kasus geringes Kompensationspotenzial für fehlende Numerus-Vorkommen auf Types-Ebene ergibt.

Numerus	**Auftreten**	**Entspricht**
Singular	53.213	67,4%
Plural	23.503	29,8%
Singular und Plural	2.204	2,8%
Singular, Plural und *	31	0,04%
Nur *	29	0,04%
Plural und *	9	0,01%
Singular und *	8	0,01%
Summe	78.997	100%

Tabelle 4.7.: Auftreten Numerus Substantive, Types

4.2. Adjektivdeklination

Die Flexion der Adjektive erfolgt in den gleichen Kategorien und mit der gleichen Menge möglicher Werte wie bei den Substantiven, das heißt in den Kasus Nominativ, Akkusativ, Genitiv und Dativ sowie den beiden Numeri Singular und Plural. Sie erfolgt jedoch regelmäßiger, da die Markierung beider Kategorien ausschließlich über die fünf Suffixe *e, en, er, es, em* erfolgt und keine Umlautungen oder sonstigen Stammveränderungen auftreten. Die Tabellen 4.8, 4.9 und 4.10 zeigen das Deklinationsschema in den von Eisenberg übernommenen, traditionellen Paradigmen starke, schwache und gemischte Deklination. Das Paradigma wird vom vorhergehenden Artikel oder der Abwesenheit eines solchen bestimmt. Der hohe Anteil von Synkretismen sowohl in Kasus als auch Numerus ist unmittelbar ersichtlich:

Kasus	Maskulin	Feminin	Neutrum	Plural
Nominativ	er	es	e	e
Akkusativ	en	es	e	e
Genitiv	en	en	er	er
Dativ	em	em	er	en

Tabelle 4.8.: Adjektivdeklination, stark

Kasus	Maskulin	Feminin	Neutrum	Plural
Nominativ	e	e	e	en
Akkusativ	en	e	e	en
Genitiv	en	en	en	en
Dativ	en	en	en	en

Tabelle 4.9.: Adjektivdeklination, schwach

Kasus	Maskulin	Feminin	Neutrum	Plural
Nominativ	er	es	e	en
Akkusativ	en	es	e	en
Genitiv	en	en	en	en
Dativ	en	en	en	en

Tabelle 4.10.: Adjektivdeklination, gemischt

4.2.1. Sonderfall Komparation

In Übereinstimmung mit Eisenberg (so etwa Regelmäßigkeit, Nichtexistenz morphologisch einfacher Formen von Komparativ und Superlativ, 2020:191f) wird die Adjektivkomparation in dieser Arbeit zur Flexion und nicht zur Derivation gezählt. Somit wird sie ebenfalls zum Gegenstand dieser Studie.

Die Bildung von Komparativ und Superlativ erfolgt über regelmäßige Suffigierung mit *er* (Komparativ) bzw. *(e)st* (Superlativ) mit oder ohne Stammumlautung (*rund – runder – (am) rundest(en)* vs. *groß – größer – (am) größt(en)*) oder mittels Suppletivformen: *gut – besser – (am) best(en)*, *viel – mehr – (am) meist(en)*. Da die benötigten Deklinationssuffixe an die Komparationsinfixe angehängt werden, handelt es sich um einen Agglutionationsvorgang. Im Falle adverbialer Verwendung des Superlativs beziehungsweise in Verbindung mit Kopulaverben tritt mit der Partikel *am* noch eine analytische Komponente hinzu.

In der Lemma-Spalte gesteigerter Adjektive notiert TübaDZ eine Art Rumpfstamm, so etwa *größt* bei *am größten*. Da es sich somit jeweils um eine vom Lemma des Positivs abweichende Zeichenkette handelt und und da diese Abweichung der Zeichenkette einen semantischen Unterschied markiert (etwa zwischen *schnell* - *schneller* - *am schnellsten*), der für die Themenklassifikation relevant sein kann, sollen Komparativ und Superlativ in Abgrenzung zu Eisenberg als eigenständige Lexeme betrachtet werden.

Da TübaDZ im ConLL-Format lediglich die morphosyntaktische Annotation zu Kasus, Numerus und Genus analog zu der der Substantive bereitstellt, nicht jedoch, ob es sich um Positiv, Komparativ oder Superlativ handelt, ist eine automatische empirische Erhebung dieser Verteilung nicht möglich. Dies ist über statistisches Interesse hinausgehend aus folgendem Grund erwähnenswert: Zwischen dem Komparativ eines Adjektivs ohne

Stammumlautung besteht, sofern kein Deklinationssuffix hinzutritt, Synkretismus mit drei Formen des Positivs in der starken sowie einer Form in der gemischten Deklination (etwa: *heißer Tee – Dieses Wasser ist heißer.*, *ein schöner Abend – Dieser Abend ist schöner.*

4.2.2. Korpusanalyse TübaDZ

99.659, also rund 6,2% der Tokens im Korpus, sind mit ADJA als attributive und somit deklinierte Adjektive getaggt. Die folgenden Betrachtungen lassen weitere 39.676 als ADJD getaggte, als undeklinierte Adjektive adverbial oder prädikativ verwendete Tokens vorerst außen vor. Bei den als ADJA getaggten Tokens handelt es sich um Mehrfachvorkommen von 7.448 Types, bei denen es sich um deklinierte Formen von 6.508 Lexemen handelt. Diese 6.508 Lexeme stellen 5,71% der im Korpus insgesamt vorkommenden 113.922 Lexeme. Sie treten im Schnitt in 1,14 deklinierten Formen auf. Dieser Abschnitt untersucht analog zum Abschnitt über die Substantive im Korpus die empirische Distribution der Adjektivdeklination.

4.2.2.1. Kasus

Die Funktion häufiger auftretender, deklinierter Adjektive als Teil von Nominalgruppen hat möglicherweise Auswirkungen auf die Repräsentativität der deklinierten Formen in den jeweiligen Kasus für das Lexem insgesamt: Lediglich 55,34% der Adjektive im Vergleich zu 70,5% der Substantive treten nur in einem Kasus auf, 13,26% in zwei, 11,47% in drei und 11,87% in allen vier Kasus. Somit treten zusammgefasst 23,34% der Adjektive in drei oder allen Kasus auf, im Vergleich zu 13,74% der Substantive.

Tabelle 4.12 zeigt die Verteilung der Kasus aus der Perspektive der Types:

Kasus	Auftreten	Entspricht
Nominativ	2.250	17,85%
Dativ	2.009	15,94%
Akkusativ	1.858	14,74%
Nominativ, Akkusativ, Dativ, Genitiv	1.496	11,87%
Genitiv	858	6,81%
Nominativ, Akkusativ, Dativ	836	6,63%
Nominativ, Dativ	604	4,79%
Nominativ, Akkusativ	558	4,43%
Akkusativ, Dativ	494	3,92%
*	454	3,60%
Nominativ, Akkusativ, Genitiv	248	1,97%
Nominativ, Genitiv	214	1,70%
Nominativ, Dativ, Genitiv	209	1,66%
Dativ, Genitiv	202	1,60%
Akkusativ, Genitiv	158	1,25%
Akkusativ, Dativ, Genitiv	152	1,21%
Summe	12.604	100%

Tabelle 4.11.: Auftreten Kasus Adjektive, Lexeme, > 1%

Kasus	Auftreten	Entspricht
Nominativ	5.802	25,82%
Dativ	4.307	19,17%
Akkusativ	3.881	17,27%
Nominativ, Akkusativ	2.535	11,28%
Genitiv	1.776	7,90%
Dativ, Genitiv	695	3,09%
Nominativ, Dativ, Akkusativ, Genitiv	599	2,67%
Akkusativ, Dativ	567	2,52%
Akkusativ, Dativ, Genitiv	499	2,22%
*	470	2,09%
Nominativ, Dativ	381	1,70%
Nominativ, Dativ, Genitiv	347	1,54%
Nominativ, Genitiv	256	1,14%
Summe	22.468	100,00%

Tabelle 4.12.: Auftreten Kasus Adjektive, Types, > 1%

70,16% der Adjektiv-Types gegenüber 75,9% der Substantiv-Types treten nur in einem Kasus, 20,39% in zwei, 4,66% in drei, 2,67% in allen vier Kasus auf. Weiter zusammengefasst fungieren 27,72% der Adjektivtypes (Substantive 23,7%) für mehr als einen und 9,43% (Substantive: 10,7%) für drei oder mehr Kasus.

Durch die mit stets fünf Suffixen erhöhte Formenvielfalt (im Vergleich zu den mitunter lediglich zwei Formen des Substantivs) schlägt sich die Regelmäßigkeit der Adjektivdeklination beim Kasus im vorliegenden Szenario offensichtlich nicht positiv in einem durch Synkretismen erhöhten Abdeckungsgrad nieder.

4.2.2.2. Numerus

Auch die Darstellung zum Numerus erfolgt analog zur Darstellung bei den Substantiven. Tabelle 4.13 zeigt die Anteile der Numeri aus der Lexem-Perspektive, Tabelle 4.14 aus der der Types.

Numerus (Tag)	**Auftreten**	**Entspricht**
Singular	5.814	46,13%
Singular und Plural	3.830	30,39%
Plural	2.496	19,8%
*	454	3,6%
Singular, Plural, *	5	0,04%
Singular, *	3	0,02%
Plural, *	2	0,02%
Summe	12.604	100%

Tabelle 4.13.: Auftreten Numerus Adjektive, Lexeme

Mit 30,43% der Lexeme, die in beiden möglichen Numeri auftreten, liegt deren Anteil um den Faktor 2,4 über dem der substantivischen Lexeme (12,65%). Eine mögliche Erklärung könnte ein Frequenzeffekt zugunsten weniger stark verbreiteter Substantive

als Begleitungen von Nominalgruppen analog zur Überlegung bei den Kasus sein, die dann jeweils in einem der beiden Numeri auftretende Substantive begleiten:

Numerus (Tag)	**Auftreten**	**Entspricht**
Singular	12.687	56,47%
Plural	4.863	20,84%
Singular und Plural	4.619	20,56%
*	474	2,11%
Singular und *	3	0,01%
Singular, Plural, *	2	0,01%
Summe	22.468	100%

Tabelle 4.14.: Auftreten Numerus Adjektive, Types

Ein gar um den Faktor 7,33 gegenüber den Substantiven verstärktes Auftreten von Types (20,57% gegenüber 2,84%), in denen beide Numeri zusammenfallen, findet eine Erklärung mutmaßlich in der starken Präsenz des Pluralsuffixes *en*, wie in den Tabellen 4.8, 4.9 und 4.10 zu sehen: 9 von 12 Pluralformen, darunter sämtliche der gemischten und schwachen Deklination, werden mit diesem Suffix gebildet. Es stellt gleichzeitig 3 von 12 Suffixen in der starken sowie jeweils 7 von 12 Suffixen der gemischten und schwachen Deklination, also mit 17 von 36 Zellen fast die Hälfte der Deklinationsfälle im Singular. Die Wahrscheinlichkeit, dass ein Type durch diesen *en*-Synkretismus ein Lexem in beiden Fällen repräsentieren kann, ist also erhöht.

Tabelle 4.15 schließlich rundet den Eindruck, dass die Adjektivdeklination Erschwernispotenzial für die Textklassifikation bereithält, aus der Perspektive der Suffixe ab: 92,14% der Adjektiv-Lexeme treten nur mit maximal drei beliebig kombinierten der sechs möglichen Endungen (inklusive Endungslosigkeit) auf, und eine deutlich absolute Mehrheit in lediglich einer der sechs möglichen Formen. Die verbleibenden fünf Formen können einem Klassifikator im Trainingsverfahren nicht bekannt sein, im späteren Klassifikationsvorgang jedoch potentiell jederzeit auftreten und die Klassifikation erschweren.

# Flexionsendungen	Lexeme	Entspricht
1	7.767	61,62%
2	2.527	20,05%
3	1.320	10,47%
4	643	5,10%
5	271	2,15%
6	76	0,60%
Summe	12.604	100%

Tabelle 4.15.: Auftreten Flexionsendungen Adjektive

4.3. Verben: Konjugation

Das verbale Konjugationsparadigma ist ungleich komplexer als das der Substantive und Adjektive. Als erstes Ordnungselement der Kategorienlehre sieht Eisenberg die Finitheit des Verbs, mit den möglichen Werten *infinit*, *finit* und *semifinit*. Infinit sind die Infinitive und das Partizip II. Die Kategorien des Infinitivs sind Tempus und Genus verbi. Die Kategorien der finiten Verben sind Person, Numerus, Modus, Tempus und Genus verbi. Einzige Kategorie des semifiniten Verbs, des Imperativs, ist der Numerus.

Entsprechend der Zielsetzung der Arbeit werden ausschließlich Aspekte der Konjugation betrachtet, die distinkte Wortformen erzeugen. Somit entfällt eine Betrachtung des Genus verbi, namentlich des Passiv, das im Deutschen nicht synthetisch, sondern mit Auxiliarverben und dem Partizip II gebildet wird. Diese entfällt ebenso für alle nichtsynthetischen Tempora.

4.3.1. Infinite Formen

Dieser Abschnitt betrachtet den reinen und den zu-Infinitiv sowie aufgrund seiner periphrastischen Funktionen das Partizip II, das als morphologisch eine Zwischenstellung einnehmend zwischen verbalem und adjektivischem Paradigma betrachtet wird. Zuvor

ist zu begründen, warum das aus dem Infinitiv abgeleitete, eigene Wortformen bildende Partizip I in diesem Zusammenhang nicht behandelt wird.

Das Partizip I wird im Deutschen regelmäßig aus dem Infinitiv gebildet (*sehend, arbeitend, seiend*, Beispiele aus Eisenberg (2020:211)). In dieser Arbeit wird das Partizip I in Übereinstimmung mit Eisenberg nicht als Teil des verbalen Paradigmas betrachtet, da es nicht periphrastisch verwendet werden kann und auch in TübaDZ dem STTS-Standard entsprechend als ADV *sehend* oder ADJA *sehenden* getaggt wird. Partizip I-Formen von Verben werden somit ohne gesonderte Unterscheidung im Zusammenhang mit den Adjektiven behandelt.

Die in den Abschnitten zur Deklination übliche Unterteilung in Einführung und Korpusanalyse wird in diesem Kapitel nur auf die komplexeren finiten Formen des Verbs angewandt. Bei Infinitiv und Imperativ wird sie hingegen auf eine unmittelbare Darstellung direkt neben der theoretischen Beschreibung komprimiert.

4.3.1.1. Infinitiv

Der deutsche Infinitiv endet bis auf wenige Ausnahmen (etwa *tun*) auf einer der drei möglichen Endungen *en*, *ern* oder *eln*, von denen die erstgenannte mit 3.829 Vorkommen in TübaDZ die häufigste ist. Diese Regelmäßigkeit wird zusätzlich auch Gegenstand der Betrachtung von Homographien sein. Neben dem als Zitierform des Verbs verwendeten Infinitiv Präsens Aktiv *auslachen* existiert ein Passiv jeweils des Infinitiv Präsens (*ausgelacht werden*) und Perfekt (*ausgelacht worden sein*). Alle außer dem erstgenannten Infinitiv Präsens werden in dieser Arbeit nicht weiter betrachtet, da analytisch und somit keine eigenen Wortformen erzeugend. Beim zu-Infinitiv gilt es für diese Arbeit zu unterscheiden zwischen Verben ohne und mit trennbaren Verbpartikeln. Bei Ersteren steht das *zu* einzeln (*zu kaufen*), während es bei Letzteren fix zwischen Partikel und Stamm steht und somit eine eigene Wortform erzeugt (*einkaufen* und *einzukaufen*, nicht

**zu einkaufen*). Beim reinen Infinitiv besteht ein Synkretismus zur 1. und 3. Person Plural Indikativ Aktiv Präsens sowie zum Imperativ Plural (*tragen – wir tragen – sie tragen – tragen Sie!*).

Eine Auszählung der infinitivbezogenen STTS-Tags in TübaDZ ergibt die in Tabelle 4.16 gezeigten Häufigkeiten.

Tag	Tokens	Types	Lexeme
VVINF	27.608	3.715	3.637
VAINF	5.732	11	7
VVIZU	2.910	1.082	1.068
VMINF	1.071	7	9

Tabelle 4.16.: Infinitivformen nach Tags, absteigend nach Anzahl Tokens

Auf dieser Verteilung basierend werden nur die als VVINF und VVIZU getaggten Wortformen näher betrachtet, da es sich bei den Modal- und Auxiliarverben um eine überschaubare geschlossene Menge (*dürfen, können, mögen, müssen, sollen* und *wollen* sowie *haben, sein* und *werden*) handelt.

4.3.1.2. Partizip II

Die zweite infinite Form des deutschen Verbs, das Partizip II, erfüllt eine Reihe von Funktionen in periphrastischen Verbformen: Es wirkt mit bei der Bildung der analytischen Tempora Perfekt, Plusquamperfekt und Futur II sowie am Zustands- und Vorgangspassiv für finite Verben und Infinitive, gebildet mit den Auxiliarverben *haben*, *sein* und *werden*. Eisenberg (2020:211) benennt Grenzfälle wie idiomatische Verwendungen, die in dieser Arbeit außer Acht gelassen werden. Darüber hinaus kann es adjektivische Verwendung finden und im Flexionsparadigma der Adjektive regulär dekliniert werden.

Die Bildung erfolgt durch Präfigierung des Infinitivstamms der starken beziehungsweise der schwachen Verben mit *ge-* sowie dem Suffix *-en* beziehungsweise *-t*. Das Präfix *ge-*

entfällt bei präfigierten Verben. Bei den starken Verben erfolgt zusätzlich in der Regel Ablautung. Bei mit trennbaren Partikeln präfigierten Verben wird *ge* zum Infix mit fester Position analog zu *zu* beim Infinitiv (*ein-ge-kauft*).

TübaDZ enthält 34.158 mit VVPP als Partizip II getaggte Tokens. Es handelt sich um Mehrfachvorkommen von 4.628 Types, die von 4.254 Lexemen gebildet werden.

4.3.2. Finite Verbformen

Die für die Untersuchung relevanten Kategorien des finiten Verbs sind Person, Numerus, Modus und Tempus.

Die Werte der Kategorie Person sind 1., 2. und 3. Person, die Werte des Numerus sind Singular und Plural, die Werte des Modus sind Indikativ und Konjunktiv.

Tabelle 4.17 zeigt bei der Auszählung der möglichen Werte der Kategorie Modus, Indikativ und Konjunktiv, einen Überhang 96,74% zugunsten des ersteren. Die Besprechung der Konjugationsmechanismen für die finiten Verbformen erfolgt für die Kategorien Person, Numerus und Tempus zunächst für den Indikativ, bevor sie auf einige Aspekte des Konjunktivs eingeht. Bei der sich anschließenden empirischen Aufstellung zu den als VVFIN getaggten Tokens wird nicht mehr gesondert nach Modus unterschieden.

	Präsens	**Anteil**	**Präteritum**	**Anteil**
Indikativ	50.000	66,46%	22.778	30,28%
Konjunktiv	1.729	2,29%	728	0,97%

Tabelle 4.17.: Aufteilung Modus finite Verbformen, Tokens

4.3.2.1. Imperativ

Der Imperativ erzeugt distinktive Formen nur im Singular. Die Pluralform entspricht der 2. Person Plural Indikativ Präsens (Beispiele: *seht, legt, segelt*, Eisenberg (2020:212)). Die Singularform entspricht bei den schwachen und den meisten starken Verben der 2. Person Singular Indikativ Präsens ohne die Personalendung *(s)t*. Vokalanhebungen starker Verben werden übernommen (*lies, sieh, wirf*). Abweichungen gelten für umlautende starke Verben wie *fährst* und *brätst*: So heißt es *fahr(e)* statt **fähr* und etwa *brat(e)* statt **brät*.

TübaDZ enthält nur 16 Tokens, die als VAIMP, also Auxiliarverb-Imperative, getaggt sind. Diese verteilen sich auf nur 5 Types, die von 3 Lexemen (*haben, sein, werden*) gebildet werden. Angesichts dieser geringen quantitativen Relevanz werden die Imperative der Auxiliarverben nicht weiteer untersucht.

Den Tag VVIMP tragen 490 Tokens, 216 Types und 195 zugrundeliegende Lexeme. Das Verhältnis Singular zu Plural bei den Tokens ist 345 zu 145. Da es sich nur bei ersteren um eigenständige Formen handelt und ihnen lediglich 133 Lexeme zugrunde liegen, werden diese ebenfalls nicht weitergehend behandelt.

4.3.2.2. Die Konjugation im Indikativ

Die folgenden Tabellen illustrieren das Konjugationsparadigma nach Eisenberg aus der Perspektive von Person und Numerus. Tabelle 4.18 zeigt das Paradigma für starke und schwache Verben im Präsens, die Tabellen 4.19 und 4.20 die Konjugation der schwachen respektive starken Verben im Präteritum.

Starke wie schwache Verben werden im Indikativ Präsens mit je vier Suffixen konjugiert.

		Singular	Plural
1.	leg	(e)	en
2.		st	t
3.		t	en

Tabelle 4.18.: Konjugation Indikativ, Präsens

		Singular	Plural
1.	legt	e	en
2.		est	et
3.		e	en

Tabelle 4.19.: Konjugation Indikativ, schwach, Präteritum

		Singular	Plural
1.	rief	-	en
2.		st	t
3.		-	en

Tabelle 4.20.: Konjugation Indikativ, stark, Präteritum

Während das schwache Verb auch im Präteritum mit vier Suffixen regelmäßig konjugiert wird, werden die Formen des starken Verbs hier mit drei Suffixen und zusätzlich mit Ablautung gebildet.

In diesen nach Numerus und Person gegliederten Teilparadigmen finden sich somit folgende Synkretismen, die analog zu den Betrachtungen bei der Deklination möglicherweise die Formenvielfalt in einem Klassifikationsmodell entschärfen können: zwei doppelt belegte Suffixe (*t* und *en* in der 3. Singular und 2. Plural respektive 1. und 3. Plural) für die Konjugation im Präsens; ebenfalls zwei doppelt belegte Suffixe für 1. und 3. je Singular und Plural beim Präteritum der schwachen Verben sowie analog weitere zwei doppelt belegte Suffixe an gleicher Stelle des Präteritums der starken Verben. Diese Synkretismen gelten jeweils nur innerhalb eines der drei Teilparadigmen, da in jedem ein eigener Stamm (im Beispiel: *legt* statt *leg*, *rief* statt *ruf*) zugrunde liegt.

4.3.2.3. Konjunktiv

Tabelle 4.21 (Eisenberg(2020:203)) stellt einige geläufige starke Verben in Indikativ und Konjunktiv gegenüber.

a.	**Indikativ**	**Konjunktiv**	**b.**	**Indikativ**	**Konjunktiv**
	trag(e)	trage		ruf(e)	rufe
	trägst	tragest		rufst	rufest
	trägt	trage		ruft	rufe
	tragen	tragen		rufen	rufen
	tragt	traget		ruft	rufet
c.	**Indikativ**	**Konjunktiv**	**d.**	**Indikativ**	**Konjunktiv**
	rat(e)	rate		reit(e)	reite
	rätst	ratest		reitest	reitest
	rät	rate		reitet	reite
	raten	raten		reiten	reiten
	ratet	ratet		reitet	reitet

Tabelle 4.21.: Konjunktiv im Vergleich zu Indikativ, Präsens, stark

Der Konjunktiv ist bei den starken Verben unterschiedlich gut markiert: Die Klassen um a., z.B. *tragen* und b., z.B. *rufen* erzeugen drei dinstinkte Formen von fünf möglichen im Paradigma (die 1. Singular Plural ist stets formgleich zur 3. Plural). Die Klasse um c., z.B. *raten* erzeugt zwei distinkte Formen und die von d., z.B. *reiten* vertretene Klasse lediglich eine obligatorisch sowie eine weitere fakultativ distinkte Form.

Die Bildung des Konjunktiv Präteritum der starken Verben erfolgt in verschiedenen Klassen mit Umlautung und Vokalanhebung, die hier angesichts der vorstehend postulierten geringen quantitativen Relevanz von nur 0,97% am Aufkommen der finiten Verbformen nicht behandelt werden. Die derart gebildeten Stämme werden regelmäßig mit den in Tabelle 4.22 präsentierten Personalendungen suffigiert.

Aufgrund der fehlenden Umlautungen erfolgt die Bildung der Konjunktivformen des Präsens analog zum Indikativ ausschließlich über Suffigierung. Der Konjunktiv Präsens wird bei den schwachen Verben nur in der 3. Person durchgängig mit einer eigenen Form gebildet (*er prüft, er prüfe*), während das Konjunktivsuffix *est* der 2. Person Singular nur bei vokalischem Auslaut zur Bildung einer dinstinkten Konjunktivform führt (*du redest, du redest vs. du malst, du malest*). Das Präteritum des Konjunktivs der schwachen Verben ist vollständig formgleich mit den jeweiligen Indikativformen. Tabelle 4.22 stellt das starke Verb *rufen* im Konjunktiv Präsens und Präteritum dem schwachen Verb *prüfen* gegenüber.

Person	**Präsens ruf**	**Präteritum rief**	**Präsens prüf**	**Präteritum prüft**
1. und 3. Singular	e	e	e	e
2. Singular	est	est	est	est
1. und 3. Plural	en	en	en	en
2. Plural	et	et	et	et

Tabelle 4.22.: Vergleich Konjunktiv Präs und Prät, stark und schwach

Offensichtlich sind die Personalsuffixe für schwache wie starke Verben vollständig gleich, lediglich von der Person abhängig (erneut inklusive Synkretismen zwischen 1. und 3. Person Plural wie im Indikativ sowie nun zusätzlich 1. und 3. Singular) und somit unabhängig vom Tempus, das über den jeweiligen Stamm markiert wird.

Zusammengefasst weist der Konjunktiv eine theoretisch nicht zu vernachlässigende Formenvielfalt zumindest der starken Verben auf. Vom geringen Anteil des Konjunktivs an den finiten Verben von 3,26% selbst im Register Zeitungsnachrichten müssen allerdings noch ein nicht erhobener Anteil schwacher Verben ohne distinkte Formen und die Synkretismen der übrigbleibenden Formen abgezogen werden. Aufgrund dieser geringen quantitativen Relevanz werden Indikativ und Konjunktiv daher zusammen unter der Statistik zu den VVFIN-Verbformen behandelt.

4.3.3. Korpusanalyse TübaDZ

Eine Auszählung der Verben im Korpus nach STTS-Tags ergibt folgende Verteilung von insgesamt 222.215 Tokens mit einem verbbezogenen Tag im Korpus:

Tag	Tokens	Types	Lexeme
VVFIN	75.235	10.662	5.502
VAFIN	55.349	119	36
VVPP	34.158	4.628	4.254
VVINF	27.608	3.715	3.637
VMFIN	17.175	114	42
VAINF	5.732	11	7
VVIZU	2.910	1.082	1.068
VAPP	2.454	7	8
VMINF	1.071	7	9
VVIMP	490	216	195
VAIMP	16	5	3
VMPP	15	4	3

Tabelle 4.23.: Verben nach Tags, absteigend nach Anzahl Tokens

Die vier Tags VVFIN, VVPP, VVINF und VVIZU stellen zusammen mit 139.811 Tokens 62,9% der auftretenden Verbformen. Zählt man die 75.524 Tokens mit VAFIN und VMFIN hinzu, ergibt sich mit 215.335 Tokens bereits ein Anteil von 96,9%. Bei der Unterscheidung zwischen Voll-, Modal- und Auxiliarverben handelt es sich jeweils um eine lediglich semantische, da diese Verben ansonsten den gleichen Konjugationsregeln unterliegen wie die Vollverben.

Die folgenden Abschnitte stellen die Situation im Korpus unter den Gesichtspunkten der Kategorien Person, Numerus und Tempus in ähnlicher Weise wie zur Deklination der Substantive und Adjektive dar:

4.3.3.1. Person

Person(en)	**Auftreten**	**Entspricht**
Nur 3.	4.478	81,39%
1. und 3.	768	13,96%
Nur 1.	109	1,98%
1.,2.,3.	91	1,65%
2. und 3.	36	0,65%
Nur 2.	20	0,36%
Summe	5.502	100,00%

Tabelle 4.24.: Auftreten Person Verben, Lexeme.

Neben der unmittelbar ersichtlichen Dominanz der 3. Person von 81,39%, die in der Deklination keine Parallele unter den möglichen Auswahlwerten Kasus und Numerus hat, sind zwei weitere sekundäre Kennzahlen zu bemerken: Die Zusammenziehung der beiden nachfolgend platzierten Werte ergibt eine dominante Kategorie *3. Person, 1. Person oder beide* mit einem Anteil von 97,33%. Der verbleibende Anteil mit Beteiligung der 2. Person ist mit 2,67% dementsprechend gering, der ausschließlich in der 2. Person vorkommender Lexeme bedeutungslos.

Person(en)	Auftreten	Entspricht
Nur 3.	9.143	85,75%
1. und 3.	878	8,23%
Nur 1.	457	4,29%
Nur 2.	137	1,28%
2. und 3.	45	0,42%
1., 2. und 3.	2	0,02%
Summe	10662	100

Tabelle 4.25.: Auftreten Person Verben, Types

Aus der Perspektive der Types zeigt sich die Gruppe *1. oder 3. Person oder beide* mit 98,27% noch dominanter. Der laut Konjugationsparadigma einzig mögliche, Tabelle 4.18 zu entnehmende Synkretismus zwischen 3. Person Singular und 2. Person Plural, nämlich im Indikativ Präsens über das gemeinsame Suffix *t*, lässt den äußerst geringen Anteil der diese beiden Personen gleichzeitig vertretenden Types plausibel erscheinen. Deutlich höher ist aufgrund der aus den Tabellen 4.18 bis 4.22 ersichtlichen, zahlreichen Synkretismen die Wahrscheinlichkeit für gleichzeitige Repräsentanz von 1. und 3. Person über Numerus, Modus und Tempus hinweg, mit dem deutlich höheren Wert repräsentierender Types von 8,23%.

4.3.3.2. Numerus

Numerus	Auftreten	Entspricht
Plural und Singular	2.397	43,57%
Nur Singular	2.205	40,08%
Nur Plural	900	16,36%
Summe	5.502	100

Tabelle 4.26.: Auftreten Numerus Verben, Lexeme

Beim Numerus zeigt sich bei der gleichen Anzahl möglicher Werte dieser Kategorie bei den Verben eine völlig andere Verteilung als bei den Substantiven: Während bei letz-

teren lediglich 12,6% der Lexeme in beiden Numeri auftreten, ist dies bei den Verben für eine relative Mehrheit der Fall. Unter den beiden verbleibenden Gruppen der Verben, die nur in einem Numerus auftreten, ist das Verhältnis deutlich ausgeglichener. Die Frage, ob es sich hier um einen Effekt von Domäne und Register handelt und dieser seine Grundlage möglicherweise in einer kleineren Gruppe in beiden Numeri auftretender semantisch breiterer Verben oder häufig verwendete Redewendungen in Nachrichtentexten hat, scheint bei einem Blick in die Häufigkeitenliste der Verben (*geben, sagen, gehen, kommen*) nicht uninteressant, ist jedoch als redaktionelle Frage von Semantik in einer morphologisch-quantitativen Arbeit nicht zu vertiefen.

Numerus	**Auftreten**	**Entspricht**
Nur Singular	6.545	61,39%
Nur Plural	3.946	37,01%
Plural und Singular	171	1,60%
Summe	10.662	100,00%

Tabelle 4.27.: Auftreten Numerus Verben, Types

Eine der bei den Substantiven ähnliche Aufteilung ergibt sich hingegen aus der Perspektive der Types, mit ebenso deutlichem Überhang des Singulars. Der sehr geringe Anteil an Types, die gleichzeitig Singular und Plural repräsentieren, erklärt sich aus demselben, einzigen möglichen Synkretismus, der auch das Zusammenfallen der Person in einem Type ermöglicht, dem des Suffixex *t* der 3. Person Singular Präsens und 2. Person Plural Präsens im Indikativ (z.B. *zwingt*).

4.3.3.3. Tempus

Die Verteilung der Tempora lässt einen Einfluss des Registers Zeitungssprache möglich erscheinen: 86,33% der Lexeme treten im Präsens oder im Präsens und zusätzlich im Präteritum auf, wobei der Anteil der Lexeme in beiden Tempora eine relative Mehrheit nur knapp zugunsten des ausschließlichen Präsens verpasst.

Numerus	Auftreten	Entspricht
Präsens	2.382	43,29%
Präsens und Präteritum	2.368	43,04%
Präteritum	751	13,65%
Präsens, Präteritum und *	1	0,02%
Summe	5.502	100,00%

Tabelle 4.28.: Auftreten Tempus Verben, Lexeme

Numerus	Auftreten	Entspricht
Präsens	6.832	64,08%
Präteritum	3.805	35,69%
Präsens und Präteritum	24	0,23%
*	1	0,01%
Summe	10.662	100,00%

Tabelle 4.29.: Auftreten Tempus Verben, Types

Die frappierende Abweichung zwischen dem Anteil von rund 43% in beiden Tempora auftretender Lexeme und der fast vollständigen Abwesenheit von Types, die beide Tempora abbilden können, erklärt sich aus dem Umstand, dass lediglich innerhalb oder unter Beteiligung des Konjunktivs beide Tempora in einer Form zusammenfallen können (*ziehen, unterschieden, stünden*).

4.4. Verben: Trennbare Verbpartikeln

Verbpräfixe und -partikeln erzeugen neue Verben aus einem Basisverb, die sich teilweise erheblich semantisch von diesem unterscheiden. Zur geschlossenen Menge der nicht trennbaren Verbpräfixe zählen *be-, ent-, er-, ge-, hinter-, ver-, zer-, a-, de(s)-, dis-, miss-, im-, in-, non-, re-, un-*. Verbpartikeln unterscheiden sich von den Verbpräfixen dadurch, dass sie den Wortakzent auf sich ziehen, in bestimmten syntaktischen Konstellationen vom Verb zu trennen sind und im zu-Infinitiv und den Partizipien (siehe Unterabschnitt

4.3.1.1) vom Stamm durch Infixe getrennt werden. Die syntaktische Trennung der Partikeln vom Finitum ist in jeder Satzform außer bei Verbendstellung obligatorisch (etwa *Ich kehre um*, *Morgen kaufen wir ein*). Die Trennung kann dabei durch beliebig viele und lange Satzglieder erfolgen (Eisenberg (2020:265)). Prinzipiell kann neben einer festen Menge von speziellen Verbpartikeln, die Homographien zu Präpositionen aufweisen können, jede Wortform aus den offenen Wortklassen als trennbare Verbpartikel fungieren. Des Weiteren existiert eine geringe Anzahl von Präfixen, die sowohl trennbar als auch nicht trennbar sein können, etwa *wider, wieder, fehl*. Als Beispiel für die Unterscheidung zwischen nicht trennbaren Präfixen und trennbaren Partikeln dient die Modifikation des Basisverbs *rufen* etwa durch die trennbaren Partikeln *ab, an, auf* und *zu* (*ab-rufen*, *an-rufen*, *auf-rufen*, *zu-rufen*) sowie durch das nicht trennbare Präfix be (*be-rufen*). Auch wenn argumentiert werden kann, dass es sich bei der Modifikation von Verben durch trennbare Präfixe oder Partikeln um Komposition handelt, wird das Phänomen der trennbaren Verbpartikeln in dieser Arbeit Eisenbergs Argumentation folgend als Flexionsphänomen betrachtet, da die Trennung von Partikel und Verbstamm von Konjugationsvorgängen ausgelöst wird (Infinitiv vs. finite Formen).

Dieses Phänomen ist von Relevanz für die automatische Textklassifikation, da ein separat stehender Verbstamm als Finitum mit dem zugrundeliegenden Basislexem verwechselt werden kann: Für ein Merkmalsauswahlverfahren oder einen Klassifikator ist der Zusammenhang zwischen getrennter Partikel und Finitum in einem wortformenbasierten Unigrammodell nicht ohne Weiteres ersichtlich. Das separat stehende Finitum ist vollständig homonym im Sinne von Lyons' Homonymiedefinition (Lyons (1990)), siehe Unterkapitel 4.5). Bei den trennbaren Verbpartikeln und ihren Auswirkungen handelt es sich mithin um ein Phänomen, das als außer zur Konjugation zugehörig auch als Aspekte der Homonymie aufweisend aufgefasst werden soll. Die Problematik der Homonymie und ihrer häufigsten Ausprägung, der Homographie, wird im folgenden Abschnitt separat genauer betrachtet.

TübaDZ enthält 178 als Präposition PREP getaggte Types und 181 als getrennte Verbpartikeln PTKVZ getaggte Types. Von diesen treten die folgenden 23 Types mit beiden Tags auf: *ab, an, auf, aus, bei, durch, entgegen, entlang, gegenüber, gleich, inne, mit, nach, nah, nahe, statt, über, um, unter, vor, wider, zu, zuwider*. Diese Untersuchung beschränkt sich aus zwei Gründen auf diese spezielle geschlossene Menge trennbarer Verbpartikeln mit Homonymien zu Präpositionen: Zum einen wird in dieser Arbeit die vereinfachende Annahme getroffen, dass Wortformen der offenen Wortklassen als getrennte Verbpartikeln separat stehend noch genügend semantische Distinktionskraft besitzen, um gegebenenfalls als selbständiges Merkmal Eingang in den Merkmalsraum zu finden (Beispiele nach Eisenberg (2020:277): *brustschwimmen, worthalten, schwarzärgern, kennenlernen, liegenlassen*). Zum anderen treten die doppelt als PTKVZ und PREP getaggten Partikeln ungleich häufiger, und, wie in der folgenden empirischen Analyse zu sehen, in unsystematischen und uneindeutigen Verteilungen zwischen beiden Wortklassen auf. Dies lässt sie als statistisch anspruchsvolles Phänomen erscheinen.

4.4.1. Korpusanalyse TübaDZ

Tabelle 4.30 präsentiert die 30 in TübaDZ am häufigsten mit trennbaren Partikeln präfigierten Verben. Der Präfigierungsgrad schwankt bereits in dieser kleinen Stichprobe zwischen rund 12% und über 96%, wobei eine gelegentlich unterbrochene Tendenz zur häufigeren Präfigierung mit abnehmender Frequenz eines Verbs zu beobachten ist. Möglicherweise können die häufigeren Verben als semantisch „breiter“ aufgefasst werden (*lassen, geben, machen, liegen* etc.). In jedem Fall zeigt sich eine große Bandbreite von Präfigierungsanteilen, die für einen Klassifikator auf Wortebene offensichtlich nicht antizipierbar sind.

Tabelle 4.31 zeigt die Relevanz der Fragestellung der trennbaren Verbpartikeln auf Textebene: Mehr als die Hälfte der Texte enthält mindestens zwei Partikelverben mit Tren-

Type	Anteil mit Partikel	# mit	# ohne	Gesamt
tauchen	96,72%	118	4	122
lehnen	93,38%	141	10	151
kündigen	91,33%	137	13	150
weisen	89,85%	177	20	197
teilen	80,53%	182	44	226
treten	75,00%	252	84	336
schließen	70,19%	113	48	161
legen	68,11%	173	81	254
werfen	68,05%	164	77	241
schlagen	66,92%	176	87	263
stellen	64,71%	497	271	768
setzen	59,86%	343	230	573
nehmen	59,50%	404	275	679
rufen	55,91%	123	97	220
fallen	49,88%	208	209	417
ziehen	49,57%	228	232	460
bieten	47,94%	128	139	267
sehen	39,09%	428	667	1.095
laufen	37,24%	143	241	384
gehen	37,16%	826	1.397	2.223
kommen	33,98%	737	1.432	2.169
halten	31,11%	242	536	778
finden	30,88%	264	591	855
führen	30,68%	135	305	440
bringen	27,51%	137	361	498
stehen	21,52%	355	1.295	1.650
liegen	17,92%	153	701	854
machen	16,29%	229	1.177	1.406
geben	13,84%	359	2.234	2.593
lassen	12,04%	157	1.147	1.304

Tabelle 4.30.: 30 Verben mit höchstem Präfigierungsgrad, absteigend nach Präfigierungsanteil

nung von Finitum und Partikel, ein Viertel wenigstens vier und immerhin noch 3,87% sogar zehn oder mehr solcher Verben. Nur eine deutliche Minderheit von 28,62% der Dokumente im Korpus ist überhaupt nicht betroffen. Beim speziellen Aspekt der trennbaren Verbpartikeln handelt es sich angesichts dieser Zahlen um einen weiteren Untersuchungsgegenstand als ebenso ubiquitäres empirisches Phänomen wie Deklination und Konjugation.

Anzahl Texte	Anzahl PTKVZ	Texte kumuliert	Anteil kumuliert
1.043	0	3.644	100,00%
757	1	2.601	71,38%
560	2	1.844	50,60%
357	3	1.284	35,24%
290	4	927	25,44%
192	5	637	17,48%
112	6	445	12,21%
77	7	333	9,14%
60	8	256	7,03%
55	9	196	5,38%
40	10	141	3,87%
17	11	101	2,77%
17	12	84	2,31%
16	13	67	1,84%
14	14	51	1,40%

Tabelle 4.31.: Anteile Texte mit getrennten Verbpartikeln

4.5. Homonymie und Homographie

In einem Bag-of-Words-Klassifikationsmodell besteht neben der Möglichkeit der Unterschätzung der Bedeutung eines Lexems durch die bisher betrachteten Flexionsphänomene auch die Möglichkeit einer Über- oder Unterschätzung von Frequenzen durch Polysemie und Homonymie. Diese Untersuchung übernimmt die Definitionen beider Konzepte von Lyons (1990:136ff): Danach handelt es sich bei einem Polysem um ein mehrdeutiges

Lexem, repräsentiert in einer Wortform, wohingegen es sich bei einem Homonym um eine Wortform handelt, in der mehrere Lexeme zusammenfallen. Homonymie kann partiell oder vollständig ausgeprägt sein. Partielle Homonymie beschränkt sich auf Homophonie, also identische Aussprache bei unterschiedlicher Schreibung, oder auf Homographie, also identische Schreibweise bei unterschiedlicher Aussprache. Homographie und Homophonie können unabhängig voneinander oder gleichzeitig in einer Wortform auftreten. Fallen beide Aspekte zusammen, spricht man von vollständiger Homonymie. Maßgebliches Kriterium in dieser Untersuchung ist aufgrund der Beschränkung auf die Textebene Homographie, so dass auch partiell homonyme Wortformen wie beispielsweise *modern* (Adjektiv vs. Verb, keine Homophonie) zum Untersuchungsgegenstand gehören.

Lexikalische Polysemie, in einem Bag-of-words-Klassifikationsmodell also das denkbare Zusammenfallen einer klassifikationsrelevanten Bedeutung eines Lexems mit irrelevanten zusätzlichen Bedeutungen in einer Zeichenkette, wird im Rahmen dieser Studie als semantische Frage nicht behandelt. Es wird überdies davon ausgegangen, dass sie durch die Heranziehung einer größeren Anzahl von Merkmalen zur Klassifikation in gewissem Ausmaß ohnehin entschärft wird. Diese Arbeit beschäftigt sich ausschließlich mit den Einflüssen speziell wortklassenübergreifender Homonymie und, durch die Beschränkung auf schriftliche Dokumente, somit wortklassenübergreifender Homographie. Der Umstand, dass die überwiegende Mehrheit der hier empirisch betrachteten Homographen offensichtlich auch homophon und somit vollständig homonym ist, ist für die Untersuchung irrelevant.

Wie die folgende empirische Analyse aufzeigt, weisen das lemmatisierte Korpus und die Liste der exportierten Lemmata aus Wiktionary (siehe Unterkapitel 3.2) einen weitaus geringeren Grad an Homographie auf als die flektierten Wortlisten und Originaldokumente. Wortklassenübergreifende Homographie im Sinne dieser Untersuchung liegt vor, wenn eine Wortform bzw. ein Lemma in mehr als einer der Listen der drei offenen Wortklassen zu finden ist.

Dies ist bei den Wortformenlisten bei 2.208 Substantiven der Fall, die auch ein Verb sein können, bei 1.043 Substantiven, die auch in der Adjektivliste zu finden sind, sowie bei 964 Adjektiven, die auch in der Verbliste vorkommen. Die Listen der homographen flektierten Wortformen enthalten somit 4.215 Wortformen (3.698 Wortformen abzüglich Überschneidungen).

Die wesentlich geringeren Zahlen zu den Listen der Lemmata zeigen, dass wortklassenübergreifende Homographie in potenziell relevantem Ausmaß von Flexionsvorgängen erzeugt wird: Lediglich 372 der Lemmata sind in den beiden Listen Substantive und Adjektive anzutreffen, 247 Substantive sind homograph zu einem Verb und zu vernachlässigende 39 Adjektiv-Lemmata sind auch in der Verbliste anzutreffen. Eine Analyse der häufigsten Bi- und Trigrammendungen der Homographen bestätigt diese Vermutung im Abschnitt zur Empirie. Vor der Darstellung der empirischen Lage in TübaDZ ist abschließend zu betonen, dass das Auftreten in zwei oder mehr Listen das einzige Kriterium für die Behandlung als wortklassenübergreifende Homographie begründet. Die Bandbreite der semantischen Nähe eines solchen Homographen zu seinem Pendant in der oder den Nachbarlisten ist groß: Von einfachen Verbalsubstantivierungen wie *essen - Essen* (Lemmata) über etymologisch begründbare Verwandtschaften wie *die wüste Party - die Wüste Gobi* (dekliniert) bis zu vollkommen zufälligen Übereinstimmungen wie *modern* (Adjektiv) - *modern* (Verb) (Lemmata). In Ermangelung einer objektiven, domänenunabhängigen Metrik für semantische Nähe auf Wortebene genügt die im Folgenden quantifizierte Feststellung, dass eine Reihe von semantisch in welchem Grad auch immer abweichenden Homographen im Korpus zu finden ist und den Klassifikationsvorgang in zu untersuchendem Ausmaß beeinflusst. Wie in Kapitel 3 erläutert, soll das Auftreten eines Lexems in mehreren Wortlisten von Wiktionary aufgrund des freien Editionsprinzips und der etablierten Qualität des offenen Wörterbuchs als hinreichende Indikation für eine Relevanz als Homograph im allgemeinen Sprachgebrauch gesehen werden.

4.5.1. Korpusanalyse TübaDZ

Die Tabellen 4.32 und 4.33 zeigen die Anzahl von Homographien in den Vollformenlisten und den Lemmatalisten aus Wiktionary inklusive Types-to-Tokens-Verhältnis.

Klasse	**Tokens**	**Types**	**Verhältnis**
Verben	37.074	1.692	1:21,91
Substantive	30.601	1.472	1:20,79
Adjektive	16.872	900	1:18,75
Summe	**84.547**	**4.064**	**1:20,80**

Tabelle 4.32.: Homographen nach Wortklassen, Wortformen

Hiernach sind 16,68% der 222.215 als Verben getaggten, 8,95% der 341.936 als Substantive getaggten und 7,09% der 238.034 als Adjektive getaggten Tokens betroffen. Aufklärung über die Ursache des auffällig hohen Homographie-Anteils an den konjugierten Verben erbringt ein Blick auf die häufigsten Vertreter dieser Kategorie:

Häufigkeit	**Type**	**Primär-Tag**
3.227	haben	VAFIN
1.761	habe	VAFIN
1.692	sein	VAINF
1.467	soll	VMFIN
975	können	VMFIN
855	haben	VAINF
767	wollen	VMFIN
697	würde	VAFIN
607	macht	VVFIN
559	können	VMINF
12.607		

Tabelle 4.33.: Top 10 Verbhomographen, Vollformen

34% der Verb-Homographien betreffen ausschließlich die häufigsten zehn Types dieser Liste, bis auf *macht* ausschließlich Auxiliar- und Modalverben. Da rechtfertigungsbedürftige redaktionelle Eingriffe in Merkmalsauswahl und Klassifikationsverfahren in dieser

Arbeit grundsätzlich unterbleiben, werden diese Types in der Liste der Homographien belassen und ihnen ein entsprechend geringer Einfluss auf Klassifikationsvorgänge unterstellt:

Klasse	**Tokens**	**Types**	**Verhältnis**
Verben	44.363	573	1:77,42
Substantive	11.661	319	1:36,55
Adjektive	7.517	324	1:23,20
Summe	**63.541**	**1.216**	**1:52,25**

Tabelle 4.34.: Homographen nach Wortklassen, Lemmata

Laut Tabelle 4.33 sind 19,96% der 222.215 als Verben getaggten, 3,41% der 341.936 als Substantive getaggten und 3,16% der 238.034 als Adjektive getaggten lemmatisierten Tokens betroffen. Während sich der Anteil der Homographen an Substantiven und Adjektiven durch den Wegfall wortklassenübergeifender Flexionssuffixe wie *en* mehr als halbiert, steigt der Anteil der homographen Verben sichtbar, so dass sich auch hier ein Blick auf die zugehörige Häufigkeitenliste lohnt.

Häufigkeit	**Type**	**Primär-Tag**
17.717	sein	VAFIN
3.047	haben	VAFIN
1.718	sagen	VVFIN
1.397	gehen	VVFIN
1.286	sein	VAINF
896	wissen	VVFIN
669	sehen	VVFIN
484	wollen	VMFIN
450	sprechen	VVFIN
449	sein	VAPP
28.113		

Tabelle 4.35.: Top 10 Verbhomographen, Lemmata

Offensichtlich verzerrt alleine die Homographie des Auxiliarverbs *sein* in seiner Grundform mit seiner Verbalsubstantivierung (*Sein oder Nichtsein*) den Anteil der homogra-

phen Verb-Lemmata so stark, dass dieser bei Nichtberücksichtigung statt 19,96% eigentlich nur 26.646 Tokens, entsprechend 11,99%, beträgt.

Die Mutmaßung, Suffigierung mit wortklassenübergreifend häufig verwendeten Flexionsmorphemen habe einen relevanten Anteil am deutlich größeren Anteil von Homographen unter den deklinierten und konjugierten Wortformen als in der Liste homographer Lemmata, kann durch Auszählung der Bi- und Trigramm-Endungen bestätigt werden: Die 10 häufigsten Bigramme stellen mit 2.661 Formen bereits die absolute Mehrheit der nichtlemmatisierten Homographen. Mit dem überproportional starken *-en* ist auf dem ersten Platz das häufigste Deklinations- und Konjugationsmorphem, verantwortlich für eine große Zahl der Pluralbildungen in allen drei Wortklassen, zu finden. Die Drittplatzierung von *-er* resultiert aus dessen Verwendung etwa für substantivische Pluralbildung, starke Adjektivdeklination und Komparation. Fünf der sonstigen sieben Substantive enden auf der ebenfalls häufigen Endung *-e* (Schwa) in Verbindung mit typischen Endkonsonanten diverser Adjektive (*g, s, t*), die gleichzeitig eine starke Präsenz in der Konjugation zeigt.

Häufigkeit	**Bigramm**
1.456	en
340	te
184	er
110	rn
106	le
101	ge
101	rt
95	ln
88	he
80	se

Tabelle 4.36.: Die 10 häufigsten Endungen von Homographen in TübaDZ, Bigramme, unlemmatisiert

Auch wenn die zehn häufigsten Trigramme der nichtlemmatisierten Homographen mit insgesamt 1.292 Vertretern keine absolute Mehrheit erreichen wie die Bigramme, stellen

sie doch auch bereits ein knappes Drittel und ebenfalls einen weit überproportionalen Anteil möglicher Trigramme. Sechs der zehn Trigramme enden ebenfalls auf dem soeben besprochenen wortklassenübergreifenden Suffix *en*. Zusätzlich finden sich mit *ern* und *eln* die beiden klassischen Infinitivendungstrigramme (siehe 4.2, Infinitive), die zusammen mit *en* sämtliche Infinitive und somit auch die Synkretismen zur 1. und 3. Person Plural Präsens bilden. Das Trigramm *ern* findet sich parallel häufig im Dativ Plural der Substantive.

Häufigkeit	**Trigramm**
390	ten
121	gen
109	len
108	ern
107	hen
104	sen
95	eln
92	rte
83	ken
83	ren

Tabelle 4.37.: Die 10 häufigsten Endungen von Homographen in TübaDZ, Trigramme, unlemmatisiert

Zur Einschätzung der empirischen Relevanz des Homographie-Komplexes auf Textebene präsentiert Tabelle 4.38 stark gekürzt die Anzahl der Texte im Korpus, die ein bis zehn Homographen enthalten. Aus diesen Zahlen geht hervor, dass lediglich 3,51% der Nachrichtentexte völlig frei von Homographen sind, während mehr als zwei Drittel von ihnen mindestens zehn Homographen enthalten.

Somit motivieren auf den ersten Blick zwei Argumente die Betrachtung der Homographie als Untersuchungsgenstand aufgrund des erheblichen Erschwernispotenzials für die Klassifikation: Die bereits aus der N-Gramm-Analyse sichtbaren Produktionsmechanismen bei flektierten Wortformen und die Anzahl erheblich betroffener Texte. Dabei werfen die

Anzahl Homographen	Anzahl Texte	Kumuliert	Anteil kumuliert
1	128	3.516	96,49%
2	171	3.345	91,79%
3	178	3.167	86,91%
4	139	3.028	83,10%
5	153	2.875	78,90%
6	95	2.780	76,29%
7	98	2.682	73,60%
8	84	2.598	71,30%
9	79	2.519	69,13%
10	69	2.450	67,23%

Tabelle 4.38.: Verteilung Anzahl Homographen pro Text

Listen in den Tabellen 4.33 und 4.35 allerdings die Frage auf, wie sich diese Konzentration im Merkmalsraum fortsetzen wird, ob die semantische Stärke dieses Phänomens in Form von Ambiguitäten bei Merkmalen also letztlich dem quantitativen Ersteindruck gerecht wird.

4.6. Formale Hypothesen

Die bisherigen Unterkapitel behandelten die Flexionsphänomene Deklination, Konjugation inklusive der getrennten Verbpartikeln und flexionsbedingte Homographie sowie ihre potenziellen Auswirkungen auf einen unigrammbasierten Textklassifikator. Der klassifikationsrelevante Effekt von Deklination und Konjugation besteht in der Aufspaltung eines Lexems auf mehrere Zeichenketten mit einer jeweils geringeren Häufigkeit. Der umgekehrte Effekt durch Homographie besteht in der Überschätzung der Häufigkeit eines potenziellen Merkmals aufgrund des Zusammenfallens mehrerer Lexeme in einem Homographen. Die eine bereits thematisierte Sonderstellung zwischen Flexion und Homographie einnehmenden Partikelverben bewirken in diesem Zusammenhang einen hybriden Effekt: Durch die Trennung der Partikel fallen das Finitum des erweiterten Verbs und

das des zugrundeliegenden Verbs in einer Form zusammen. Handelt es sich beim nichtpräfigierten Verb um ein klassifikationsrelevantes Lexem, wird seine Häufigkeit durch den Zusammenfall mit dem Finitum des erweiterten Verbs überschätzt. Handelt es sich hingegen beim erweiterten Verb um ein klassifikationsrelevantes Merkmal, wird umgekehrt durch die Trennung der Partikel in der konjugierten Form dessen Häufigkeit zugunsten des Basislexems unterschätzt.

Die vorstehenden Überlegungen sind nun in Zusammenhang zu den in Kapitel 2 besprochenen Schritten des Klassifikationsprozesses, Merkmalsauswahl, Dokumentenvergleich und Klassifikationsfunktion, zu setzen. Abschnitt 4.6.1 beleuchtet Einflüsse auf die Merkmalsauswahl mittels χ^2-Test, Abschnitt 4.6.2 die Beeinflussung der TF-IDF und der Kosinusähnlichkeit und Abschnitt 4.6.3 potentielle Auswirkungen auf die verschiedenen Klassifikationsverfahren.

4.6.1. Flexion und Homographie im χ^2-Auswahlverfahren

Die in Unterabschnitt 2.2.1 eingeführte Gleichung zum Merkmalsauswahlkriterium χ^2, hier wiederholt,

$$\chi^2 = \frac{(O_k - E_k)^2}{E_k} = \sum_{e_t \in 1,0} \sum_{e_c \in 1,0} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}, \tag{4.1}$$

quantifiziert die Korrelation eines Merkmals mit einer Kategorie. Implizit wurde hierbei in der Einführung die Annahme getroffen, bei einem solchem Merkmal handle es sich um eine Wortform. Basierend auf den bisherigen Überlegungen zur Aufspaltung eines Lexems in mehrere deklinierte oder konjugierte Wortformen, wobei das zugrundeliegende Lexem die eigentlich bedeutungstragende Einheit sei, soll das Auftreten eines Lexems in

einem unlemmatisierten Dokument als Summe der Vorkommen seiner flektierten Formen verstanden werden. Dies gilt für die tatsächliche Häufigkeit wie den Erwartungswert gleichermaßen. Somit beträgt der Testwert für ein Lexem, zusammengesetzt aus seinen flektierten Formen, eigentlich

$$\chi^2 = \frac{(\sum_{i=1}^{n} o_i - \sum_{i=1}^{n} e_i)^2}{\sum_{i=1}^{n} e_i} \tag{4.2}$$

wobei n die Anzahl der von diesem Lexem gebildeten Formen bezeichnet (zu deren Verteilung siehe Unterkapitel 4.1ff).

Die folgende Beispielgleichung zeigt den Effekt der Aufsummierung der Vorkommen aller flektierten Formen durch Lemmatisierung auf den χ^2-Test für das Lexem *Burg* der Deklinationsklasse 4a, das nur zwei Formen bildet:

$$\chi^2(Burg(Lexem)) = \frac{((o_1 + o_2) - (e_1 + e_2))^2}{e_1 + e_2} \tag{4.3}$$

wobei o_1 die tatsächliche Anzahl der *Burg* enthaltenden Dokumente einer bestimmten Klasse und o_2 die tatsächliche Anzahl von *Burgen* enthaltenden Dokumente bezeichnet. e_1 und e_2 bezeichnen analog die erwarteten Häufigkeiten.

Bei Einsetzen der Beispielwerte 5, 4, 2 und 1 für die genannten Variablen ergibt sich beispielsweise

$$\chi^2(Burg, Burgen) = \frac{((5+4)-(2+1))^2}{2+1} = \frac{36}{3} = 12 \tag{4.4}$$

gegenüber

$$\chi^2(Burg) = \frac{(5-2)^2}{2} = \frac{9}{2} = 4,5 \tag{4.5}$$

und

$$\chi^2(Burgen) = \frac{(4-1)^2}{1} = \frac{9}{1} = 9 \tag{4.6}$$

für die Einzelformen *Burg* und *Burgen*. Beide flektierten Formen erhalten niedrigere χ^2-Werte als das Lexem bei Addition der Einzelvorkommen. Der Umstand, dass die Summe der χ^2-Werte der einzelnen Wortformen die des Lexems bei Zusammenrechnung übersteigt, ist bedeutungslos angesichts des Umstandes, dass der Zusammenhang zwischen den beiden Zeichenketten einer den Test durchführenden Software ohne gesonderte Kennzeichnung nicht bekannt ist. Die Werte entscheiden unabhängig einzeln bewertet über die Aufnahme und Positionierung eines jeden Merkmals im Merkmalsraum.

Bei Einsetzen der Beispielwerte 5, 4, 1 und 2 für die genannten Variablen, also einem Tausch der Erwartungswerte der Einzelformen aufgrund umgekehrter Verteilung, ergibt sich hingegen

$$\chi^2(Burg, Burgen) = \frac{((5+4)-(1+2))^2}{1+2} = \frac{36}{3} = 12 \tag{4.7}$$

(also ein identischer Wert für das Lexem) gegenüber

$$\chi^2(Burg) = \frac{(5-1)^2}{1} = \frac{16}{1} = 16 \tag{4.8}$$

und

$$\chi^2(Burgen) = \frac{(4-2)^2}{2} = \frac{4}{2} = 2 \tag{4.9}$$

Die flektierte Form *Burg*, etwa als Nominativ Singular, erhält nun einen höheren χ^2-Wert als zusammengezogen mit der Pluralform im gemeinsamen Wert für das zugrundeliegende Lexem.

Bereits diese einfachen generischen Beispiele mit lediglich zwei Formen pro Lexem zeigen, dass ein Lexem nach einer Lemmatisierung und gemeinsamer Wertung aller seiner flektierten Formen sowohl einen höheren als auch einen niedrigeren χ^2-Wert erhalten kann als einzelne seiner Wortformen. Somit kann nicht pauschal vorausgesetzt werden, dass eine Lemmatisierung den χ^2-Wert erhöht. Nichtsdestrotrotz soll in Ermangelung begründeter abweichender Annahmen unabhängig von der Verteilung der einzelnen Formen eines Lexems im Klassifikationsszenario angenommmen werden, dass die Abweichungen zwischen Erwartungswert über alle Klassen und tatsächlicher Häufigkeit in einer Klasse

bei allen Formen eines Lexems in etwa gleich sind. Ist dies der Fall, schlägt die quadratisch steigende Abweichung die linear steigende Normalisierung durch Hinzufügung weiterer flektierter Formen, und die Lemmatisierung resultiert in einem höheren χ^2-Wert. Dieser Effekt wiederum fällt umso stärker aus, je mehr einzelne Formen das Lexem bilden kann. Somit sollten verbale Klassifikationsmerkmale von einer Lemmatisierung am stärksten profitieren, Substantive mit ihrem geringeren Formenreichtum am wenigsten. Umgekehrt argumentiert sollten Substantive aus einem unlemmatisierten Korpus einen signifikant höheren Anteil im Merkmalsraum im Verhältnis zu ihrem Anteil an der Menge der Types einnehmen. Dieser Effekt schwächt sich möglicherweise mit zunehmender Merkmalsraumgröße ab. Diese Annahmen werden in in Unterkapitel 5.1 experimentell überprüft.

Der Effekt von Homographie auf den χ^2-Wert eines Lexems ist ebenfalls nicht pauschal eindeutig bestimmbar: Falls vorhanden, erhöhen homographe Lesarten der Wortform die scheinbare Häufigkeit der jeweils anderen Lesart(en) in einer komplexen Interaktion, die am Beispiel der homographen Wortform *bekannte* dargestellt sei.

Die deklinierte Adjektivform *bekannte*, als stark flektierte feminine Form des Lexems *bekannt* im Singular, und die konjugierte Verbform *bekannte*, etwa als 3. Person Singular Indikativ Präteritum des Lexems *bekennen*, bilden eine gemeinsam ausgezählte Zeichenkette, deren Aufteilung zwischen den beiden Lexemen im Korpus unbekannt ist. Die Häufigkeit beider Lesarten wird nun jeweils durch die Präsenz der jeweils anderen Lesart überschätzt. Der χ^2-Wert einer homographen Wortform beträgt

$$\chi^2 = \frac{((o + s_o) - (e + s_e))^2}{e + s_e}, \tag{4.10}$$

wobei es sich bei s_o und s_e um Stör- oder Residualterme zu den beobachteten respektive erwarteten Häufigkeiten o und e handelt. Dies gilt unabhängig davon, ob es sich bei der Wortform um ein Lemma oder eine flektierte Wortform handelt. Alle vier Variablen o, e, s_o und s_e sind statistisch unabhängig voneinander: Die Lesarten der flektierten Wortformen oder Lemmata, das heißt die zugrundeliegenden Lexeme, können beliebige absolute Dokumentenfrequenzen besitzen und dabei jede beliebige Verteilung untereinander einnehmen. Die Störterme können des Weiteren statt wie in diesem Biespiel zwei ebenso drei Lexeme der offenen Wortklassen zusammengefasst modellieren, so dass s_o und s_e weiter gefasst dem Vorkommen von sämtlichen Lexemen anderer Lesart summiert entsprechen. Die Stärke dieses Verzerrungseffekts wird von zwei Faktoren beeinflusst:

– Der absoluten Dokumentenfrequenz der korrekten Lesart bei fixer Störgröße durch alternative Lesarten: Eine absolute Störgröße von 1 durch alternative Lesarten wirkt sich wie in Gleichung 4.7 erkennbar bei einer absoluten Dokumentenfrequenz von 20 wesentlich weniger stark aus als bei einer Dokumentenfrequenz von 4. Des Weiteren mindern Störgrößen mit zunehmend höheren absoluten Dokumentenfrequenzen der korrekten Lesart durch Erhöhung des Divisors weniger, als sie zu deren Anwachsen in der quadrierten Abweichung im Dividenden beitragen.

– Der Abweichung zwischen Erwartungswert und tatsächlichen Häufigkeiten der Nebenlesart, das heißt, der Aufteilung ihres Vorkommens zwischen den Störgrößen s_e und o_e: Handelt es sich um eine Gleichverteilung, das heißt, die alternative Lesart weist keine besondere Korrelation mit einer Klasse auf (bezeichnet also ein für dieses Klassifikationsszenario als Merkmal irrelevantes Merkmal), werden beide Terme gleich stark erhöht. Die absolute Abweichung und somit der Dividend bleibt gleich, durch den erhöhten Divisor sinkt jedoch der absolute χ^2-Wert. Die irrelevante zusätzliche Lesart verringert somit die Chancen des eigentlichen Merkmals auf Einzug in den Merkmalsraum. Trägt die Nebenlesart hingegen eine eigene klassifikationsrelevante Bedeutung, die sich in einer erhöhten Korrelation mit einer Klasse äußert, beeinflusst sie den χ^2-Wert in Abhängigkeit von der

Aufteilung ihres Vorkommens auf s_e und o_e auf zwei entgegengesetzte Arten: Erhöht ihr Vorkommen über den Störterm e_s überwiegend den Erwartungswert, mangels Vorkommen den Störterm der tatsächlich beobachteten Häufigkeit jedoch nur wenig, verringert sich die quadratische Abweichung zwischen Erwartungswert und tatsächlichem Wert. Diese wird des Weiteren durch einen gewachsenen Erwartungswert im Divisor weiter reduziert. Im umgekehrten Fall, einer starken Erhöhung der vermeintlichen beobachteten Häufigkeit zuungunsten des Erwartungswertes, erhöht sich die Abweichung quadratisch um den zugeschlagenen Wert der Nebenlesart, nur teilweise kompensiert durch den erhöhten Erwartungswert im Divisor. Die Gleichungen 4.11 bis 4.14 demonstrieren mit einem generischen Beispiel die Entwicklung des χ^2-Wertes eines Merkmals mit einer beobachteten absoluten Dokumentenfrequenz von 15 und einem Erwartungswert von 10, wenn ein zweites, die selbe Zeichenkette belegendes Lexem hinzugezogen wird. Die Modellierung geht von einem Szenario mit zwei Klassen, einer anfänglich absoluten Dokumentenfrequenz des beobachteten Merkmals von 20, somit eines Erwartungswertes von 10 und einer Erhöhung durch das zweite Lexem um 4 aus. Das Merkmal, das in der von ihm beeinflussten Klasse in 15 statt der zu erwartenden zehn Dokumente auftritt, erhält einen χ^2-Wert von

$$\chi^2 = \frac{(15-10)^2}{10} = \frac{25}{10} = 2,5 \tag{4.11}$$

wenn es auf Grundlage eines einzelnen Lexems gebildet wird. Tritt nun das zweite Lexem in identischer Wortform mit einer Dokumentenfrequenz von 4 hinzu, ergibt sich bei einem neuen Erwartungswert von 12

$$\chi^2 = \frac{((15+2)-(10+2))^2}{10+2} = \frac{25}{12} = 2,083 \tag{4.12}$$

also eine moderate Verringerung des χ^2-Wertes bei Gleichverteilung als irrelevantes Merkmal,

$$\chi^2 = \frac{((15+0)-(10+2))^2}{10+2} = \frac{1}{12} = 0,75 \tag{4.13}$$

also eine starke Reduktion des Wertes bei Vorkommen ausschließlich in der komplementären Klasse, sowie

$$\chi^2 = \frac{((15+4)-(10+0))^2}{10+2} = \frac{81}{12} = 6,75 \tag{4.14}$$

also eine deutliche Erhöhung bei Vorkommen ausschließlich in derselben Klasse wie das zuerst betrachtete Merkmal.

Bei positiver Differenz der Störgröße auf die tatsächliche Häufigkeit zur Störgröße auf die erwartete Häufigkeit erhöht sich somit der χ^2-Wert, bei Differenzen kleiner oder gleich Null verringert er sich. Unter der (generell bei der Verwendung von χ^2 als Merkmalsauswahlkriterium getroffenen) Annahme, dass eine Korrelation zwischen Term und Klasse einen semantischen Hintergrund hat, ist die Veränderung des χ^2-Wertes offensichtlich auch von der semantischen Nähe der in einer Wortform zusammenfallenden Lexeme beeinflusst. So sollten etwa die Lexeme *modern* (Adjektiv) und *modern* (Verb) seltener zufällig stark gemeinsam in einer Klasse auftreten als die Lexeme *betten* und *Bett* in der

Zeichenkette *betten* (bei angenommener Nichtverfügbarkeit von oder Verzicht auf Groß- und Kleinschreibung).

Gleichung 4.2 zur Bestimmung des χ^2-Wertes eines i Formen bildenden Lexems unter Zusammenziehung der Häufigkeiten seiner Einzelwerte wird nun zur Berücksichtigung beider Einflüsse, Flexion und Homographie, für den unlemmatisierten Fall erweitert auf

$$\chi^2 = \frac{((\sum_{i=1}^{n} o_i + \sum_{i=1}^{n} so_i) - (\sum_{i=1}^{n} e_i + \sum_{i=1}^{n} se_i))^2}{(\sum_{i=1}^{n} e_i + \sum_{i=1}^{n} se_i)} \tag{4.15}$$

Die Konjugation der Partikelverben vereint wie in Unterkapitel 4.4 identifiziert Aspekte beider hier vorgestellter Effekte: Ein konjugiertes Partikelverb mit erhaltener Partikel unterliegt den in Gleichung 4.2 modellierten Einflüssen der Aufspaltung auf verschiedene Terme (*einkaufe* – *einkaufst* – *einkauft*). Zusätzlich kann die Wortform mit angeschlossener Partikel potenziell bereits Homographen enthalten (beispielsweise *aufrufen* – Verb vs. Substantiv). Wird aufgrund von syntaktischen Erfordernissen die Partikel getrennt, treten zu dieser Aufspaltung und der potenziellen Homographie zwei zusätzliche Homographiequellen: Das verbleibende Finitum ist in jedem Fall wortklassenintern homograph mit seinem Basislexem, und die neue Form ist potentiell homograph mit Wortformen aus den übrigen offenen Wortklassen (Beispiel: *betten um*, *betten* als Verb oder Substantiv).

4.6.2. TF-IDF

Der beschriebene Einfluss von Flexion und Homographie auf den χ^2-Test bildet nur die erste Stufe zur Veränderung der tatsächlich für die Klassifikation herangezogenen

Merkmalsmenge: Abschnitt 2.1.2 führt das Frequenzmaß $TF - IDF$ zur Konvertierung ausgewählter Merkmale in ihre normalisierten Häufigkeiten im Dokument, gewichtet mit ihrer Verbreitung im Korpus, mit den Gleichungen

$$tf - idf_{t,d} = tf_{t,d} * idf_t \tag{4.16}$$

sowie

$$idf_t = log\frac{N}{df_t} \tag{4.17}$$

ein. Unabhängig davon, ob unlemmatisierte oder lemmatisierte Wortformen als Merkmale verwendet werden, unterliegt deren Konvertierung in $TF - IDF$-Werte erneut einer Beeinflussung durch Homographie. Aus diesem Grund ist zunächst bei der Bestimmung der tatsächlichen Termfrequenz eines Lexems eine zusätzliche Störgröße für Homographen zu addieren, so dass

$$TF(w) = TF(l) + \sum_{i=1}^{n} S_i \tag{4.18}$$

wobei S_i die absolute Häufigkeit eines Homographen dieser Wortform und i die Anzahl dieser Homographen ist.

Die beobachtete Termfrequenz einer Wortform $TF(w)$ beispielsweise *bekannte*, setzt sich zusammen aus der Häufigkeit der als Merkmal relevanten Lesart des Lexems $TF(l)$,

beispielsweise *bekennen*, sowie deren Nebenlesarten in Form von Homographen. Beträgt etwa die beobachtete absolute Häufigkeit der Wortform *bekannte* 5, davon 2-mal als konjugierte Verbform sowie je 1-mal als Substantiv *Bekannte* und 1 mal als deklinerte Form des Adjektivs *bekannt*, beträgt umgestellt

$$TF(l) = TF(w) - \sum_{i=1}^{n} S_i = 5 - (2 + 1) = 2 \tag{4.19}$$

Der Einfluss dieser Störgröße auf die inverse Dokumentenfrequenz ist komplexer, wie die Analyse der zu

$$idf_t = log \frac{N}{\sum\limits_{i=1}^{n} df_{ti}} \tag{4.20}$$

modifizierten Gleichung 2.4 zeigt: Die Dokumentenfrequenz beinhaltet nun neben Dokumenten mit der relevanten Lesart der Wortform unter Umständen weitere, stattdessen die Homographen beinhaltende Dokumente. Der Index i iteriert hier über alle denkbaren, das heißt die als Merkmal relevanten und sämtliche homographen, Lesarten. Bei fixer Korpusgröße N verringern diese den Term $\sum\limits_{i=1}^{n} / df_{ti}$ durch den erhöhten Divisor zwar linear, haben aber durch die abschließende Logarithmierung abhängig von der absoluten Termfrequenz des eigentlichen Merkmals unterschiedlich großen Einfluss: Die Stauchungswirkung der Logarithmierung ist abhängig von der Höhe der Termfrequenz, so dass dieselbe Summe hinzuaddierter Dokumentenfrequenzen der Homographen umso weniger Einfluss auf den gesamten Term ausübt, je größer er bereits für das korrekte Merkmal ist. Die Verringerung des IDF-Wertes durch Homographen unlemmatisierter Wortformen fällt also für häufige, und damit potenziell weniger relevante, Merkmale

stärker aus. Des Weiteren ist davon auszugehen, dass merkmalsrelevante Lesarten gegenüber ihren Homographen am Vorkommen der Wortform anteilig stärker ins Gewicht fallen und deren, auch linearen, Einfluss ohnehin begrenzen.

Nach der Beeinflussung der Merkmalsauswahl und des Konvertierungsprozesses zu TF-IDF zeigt sich im Folgenden, dass auch der direkte Vergleich von Dokumentvektoren den Einflüssen von Flexion unterliegt.

4.6.3. Kosinusähnlichkeit

Die in Abschnitt 2.1.2 als Gleichung 2.5 eingeführte Kosinusähnlichkeit,

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|} = \frac{\sum_{i=1}^{n} v_i u_i}{\sqrt{\sum_{i=1}^{n} (v_i)^2}\sqrt{\sum_{i=1}^{n} (u_i)^2}} \tag{4.21}$$

ermöglicht den Vergleich zweier Dokumentvektoren, deren Einträge etwa als $TF-IDF$-Termfrequenzen vorliegen. Bei der Einführung der Vektoren wurde keine Annahme getroffen, ob es sich bei den Dokumenten um unlemmatisierte oder verarbeitete, etwa lemmatisierte, Dokumente handelt. Die Untersuchung des Verhaltens der Kosinusähnlichkeit in Abhängigkeit von Flexion und Homographie soll zunächst mit einem generischen Beispiel eröffnet werden.

Die Vektoren zweier Dokumente mit dem Vokabular *{modern, moderne, kunst}* sollen in unlemmatisiertem und lemmatisiertem Zustand verglichen werden. Zwei unlemmatisierte Dokumente, repräsentiert in den Vektoren $\vec{d_1}$ und $\vec{d_2}$,

$$\vec{d_1} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad \vec{d_2} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

werden nach einer Lemmatisierung durch die Vektoren

$$\vec{d_{1l}} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad \vec{d_{2l}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

dargestellt. Die Werte der ersten beiden Stellen der Originalvektoren, die Wortformen *modern* und *moderne* repräsentierend, wurden zu einer das Lexem *modern* (Adjektiv) repräsentierenden ersten Stelle in den Vektoren $\vec{d_{1l}}$ und $\vec{d_{2l}}$ addiert. Die ehemals von der Wortform *kunst* (die in diesem Fall der Zitierform ihres Lexems entspricht) belegte dritte Stelle findet sich nun mit unverändertem Wert als das Lexem *kunst* repräsentierende Stelle des neuen Vektors wieder, da keine weiteren flektierten Formen von *kunst* gezählt werden.

Die Kosinusähnlichkeit der durch $\vec{d_1}$ und $\vec{d_1}$ dargestellten Dokumente beträgt $0,5478$ und erhöht sich durch die Lemmatisierung in die Vektoren $\vec{d_{1l}}$ und $\vec{d_{2l}}$ auf $0,7071$.

Eine höhere Kosinusähnlichkeit der lemmatisierten Vektoren gegenüber denen der unlemmatisierten Originaldokumente erscheint aus linguistischer Sicht im Kontext der inhaltlichen Textklassifikation wünschenswert: Unter der getroffenen Annahme, dass der semantische Wert eines Merkmals (in dieser Untersuchung eines Lexems) unabhängig von seiner konkreten morphologischen Erscheinung ist, repräsentieren zusammengezogene Vektorstellen der flektierten Formen eines Lexems den Inhalt eines Dokuments möglicherweise kompakter.

Bereits eine simple Modifikation nur eines Vektors im generischen Beispiel zeigt jedoch, dass eine Erhöhung der Kosinusähnlichkeit durch Lemmatisierung keineswegs generell vorausgesetzt werden kann: Der Vergleich der Vektoren $\vec{d_1}$ und $\vec{d_3}$

$$\vec{d_1} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad \vec{d_3} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \quad .$$

wobei es sich bei $\vec{d_3}$ um eine modifizierte Version des Vektors $\vec{d_2}$ mit gleichem absolutem Vorkommen des Lexems *modern* handelt, ergibt eine Kosinusähnlichkeit von $0,7303$. Da sich im modifizierten Vektor lediglich die Verteilung der aufzusummierenden Vorkommen der flektierten Formen von *modern*, also der ersten beiden Stellen, geändert hat, entspricht die lemmatisierte Version von $\vec{d_3}$ exakt dem Vektor $\vec{d_2}l$, so dass

$$\cos(\vec{d_1}l, \vec{d_3}l) = \cos(\vec{d_1}l, \vec{d_2}l) = 0,7071 \tag{4.22}$$

Die Kosinusähnlichkeit zweier Dokumentvektoren kann also bei gleicher absoluter Häufigkeit der vertretenen Lexeme, aber unterschiedlicher Verteilung auf deren flektierte Wortformen durch Lemmatisierung sowohl erhöht als auch verringert werden.

Die Auswirkungen der Lemmatisierung sind analytisch nicht pauschal lösbar, da sämtliche an der Berechnung der Kosinusähnlichkeit beteiligten Variablen untereinander unabhängig sind und auch ihre Anzahl an sich nicht festgelegt ist. Ihre Veränderung durch Lemmatisierung wird daher an dieser Stelle lediglich skizziert; diese Effekte werden im folgenden Kapitel empirisch untersucht.

– Die Dimension eines Dokumentvektors nach einer Lemmatisierung ist stets geringer oder gleich der des „unlemmatisierten“ Ausgangsvektors: Ein Lexem kann in genau einer oder mehreren flektierten Formen vorkommen. Die Anzahl der im unlemmatisierten Vektor durch diese besetzten Stellen ist aber keineswegs direkt abhängig von der morphologischen Produktivität des Lexems, sondern davon, welche flektierten Formen tatsächlich im Korpus überhaupt vorkommen, und wie viele von diesen vom Merkmalsauswahlverfahren als statistisch aussagekräftig genug selektiert wurden. Die Untergrenze für die Aufnahme in den Merkmalsraum und somit indirekt Stellen pro Lexem wiederum hängt indirekt an der Gesamtaufnahmefähigkeit des Dokumentvektors, de facto der Merkmalsraumgröße, bei der es sich um einen erlernten Parameter handelt.

– Die Dimension und die Besetzung der einzelnen Stellen beeinflusst die euklidische Länge eines Vektors. Sie ist nach Zusammenziehen mehrerer Stellen durch Lemmatisierung stets größer oder gleich der euklischen Länge des Ausgangsvektors, da stets $\sqrt{(\sum_{i=1}^{n} v_i)^2} \geq \sqrt{\sum_{i=1}^{n} (v_i)^2}$ für alle $v_i \geq 0$, eine Bedingung, die bei Dokumentenfrequenzen in jedem Fall erfüllt ist. Somit ist garantiert, dass der neue Divisor als Produkt der euklischen Längen beider Dokumentvektoren in jedem Fall größer als oder gleich dem alten ist.

– Die Veränderung des Skalarproduktes im Dividenden und die Frage, ob es den erhöhten Divisor kompensieren kann oder nicht, entscheidet über eine erhöhte oder verringerte Kosinusähnlichkeit der lemmatisierten Vektoren. Das Skalarprodukt erhöht sich umso stärker, je ungleicher die einzelnen Stellen vor der Lemmatisierung verteilt waren. Als generisches Beispiel hierzu fungiere ein Skalarprodukt zweier zweistelliger Vektoren mit unterschiedlicher Verteilung desselben Lexems, etwa

$$\vec{d_1} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \qquad \vec{d_2} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad .$$

mit dem Skalarprodukt

$$\vec{d_1} \cdot \vec{d_2} = 3 * 2 + 2 * 3 = 12 \tag{4.23}$$

im Vergleich zu

$$\vec{d_1} = \begin{pmatrix} 4 \\ 1 \end{pmatrix} \qquad \vec{d_2} = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \quad .$$

mit dem Skalarprodukt

$$\vec{d_1} \cdot \vec{d_2} = 4 * 1 + 1 * 4 = 8 \tag{4.24}$$

sowie

$$\vec{d_1} = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \qquad \vec{d_2} = \begin{pmatrix} 0 \\ 5 \end{pmatrix} \quad .$$

mit dem Skalarprodukt

$$\vec{d_1} \cdot \vec{d_2} = 5 * 0 + 0 * 5 = 0 \tag{4.25}$$

Das Skalarprodukt dieser Vektoren bei Zusammenziehung beider Stellen auf ein Lemma beträgt unabhängig von der ursprünglichen Verteilung in jedem Fall $5 * 5 = 25$. Dieser verteilungsabhängige Zuwachs kann wiederum nicht isoliert betrachtet werden, sondern nur als Teil der Aufsummierung der übrigen Dimensionen der Dokumentvektoren, also weiterer, statistisch unabhängiger Lexeme und ihrer internen Verteilungen.

4.6.4. Spezielle Einflüsse auf Klassifikationsverfahren

Nachdem in vorhergehenden Abschnitten mit der Merkmalsauswahl durch χ^2 und der Konvertierung von Wortformen in ihre $TF - IDF$-Werte zwei Arbeitsschritte identifiziert wurden, die in der konventionellen Klassifikationsarchitektur als Fehlerquellen unter dem Einfluss von Flexion in Frage kommen, werden in diesem Unterkapitel die mutmaßlichen Effekte dieser Beeinflussungen auf verschiedene Klassifikatoren besprochen. Hierbei werden lineare, unmittelbar vektorbasierte und neuronale Klassifikatoren getrennt behandelt.

4.6.4.1. Lineare Klassifikatoren

Der in Gleichung 2.17 eingeführte Typ des linearen Klassifikators

$$y = f(\vec{w} \cdot \vec{x}) = f(\sum_{i=1}^{n} w_i x_i) \tag{4.26}$$

erlernt einen Vektor von Gewichten $\vec{w}$ zu einer Reihe von Merkmalen $\vec{x}$, die gegebenenfalls linear kombiniert als Eingabe für eine Transformationsfunktion fungieren. Ein Dokumentenvektor $\vec{x}$ unterliegt dabei sowohl als Trainings- als auch Testvektor den durch die Vorverarbeitung verursachten Einflüssen. Die Belegung des Vektors durch eine bestimmte Menge von Merkmalen wurde durch das Merkmalsauswahlverfahren bestimmt und

ist im Lernprozess des Klassifikators entsprechend nicht mehr korrigierbar – durch unzutreffend niedrige $\chi^2 - Werte$ herausgefilterte, unter Umständen nützlichere Merkmale sind für die Klassifikation an dieser Stelle verloren. Kompensierbar erscheint jedoch der Effekt verzerrter $TF - IDF$-Werte an den einzelnen Stellen: Ist ein beliebiges Merkmal x_i etwa durch Homographien zu hoch bewertet, kann dieser Effekt durch Anpassung des zugehörigen Gewichts w_i kompensiert werden. Da w_i ein frei erlernbarer Parameter ist, kann er mit einem inkorrekt bewerteten Gegenstück x_i letztlich ein korrektes Produkt im Sinne der Klassifikationsleistung bilden. Die Übergabe an die Transformationsfunktion erfolgt erst nach Gewichtung und Aufsummierung aller Merkmale, so dass eine Kompensation falscher Werte an beliebig vielen Stellen durch das Erlernen angepasster Gewichte kompensierbar erscheint. Klassifikatoren, die eine lineare Kombination gewichteter Merkmale verwenden, sollten daher eine gewisse Robustheit gegenüber frequenzbasierten Fehlbewertungen ihrer Merkmale aufweisen. Unklar erscheint an dieser Stelle, in welchem Umfang durch eine abweichende Interpretation der Merkmale durch höhere oder niedrigere Gewichtung die mögliche Abwesenheit anderer wichtiger Merkmale kompensierbar wird.

4.6.4.2. Vektorbasierte Klassifikatoren

Der in Unterabschnitt 2.4.2.1 vorgestellte K-nächste-Nachbarn-Klassifikator trifft eine Klassifikationsentscheidung auf Basis der k nächsten Nachbarn eines Dokuments in dem Vektorraum, der die verwendeten Dokumente beschreibt. Zur Bestimmung der Menge dieser k Nachbarn verwendet er eine vektorbasierte Ähnlichkeitsmetrik wie die Kosinusähnlichkeit. Abschnitt 4.6.3 zeigt, dass die Kosinusähnlichkeit interagierenden und analytisch nicht auflösbaren Einflüssen von Flexion und Homographie zwischen den Merkmalen unterliegt. Diese Einflüsse können durch Lemmatisierung möglicherweise gemildert werden, auch wenn nicht pauschal vorausgesetzt werden kann, dass zwei lemmatisierte Vektoren einander ähnlicher sind als im unlemmatisierten Zustand. Bei den Stellen der

Vektoren handelt es sich um die Werte zu voneinander unabhängigen Merkmalen. Aufgrund des beschriebenen, komplexen Zusammenwirkens dieser Stellen in der Funktion, das letztlich ihre Kosinusähnlichkeit bestimmt, können verringerte Ähnlichkeiten an der einen Stelle durch stärkere Ähnlichkeiten an anderer Stelle überkompensiert werden. Dies kann die Kosinusähnlichkeit insgesamt erhöhen oder verringern. Im Zuge einer Lemmatisierung kann daher prinzipiell ein Vektor, der vorher der nächste Nachbar eines anderen Vektors war, durch einen dritten, nun ähnlicheren Vektor ersetzt werden. Dies gilt auch dann, wenn unterstellt wird, dass empirisch die Kosinusähnlichkeit von Vektoren durch Lemmatisierung in der Mehrzahl der Fälle erhöht wird: Da eine solche Erhöhung durch Ausreißer in den vektorinternen Verteilungen unterschiedlich stark pro Vektor ausfallen kann, steigt die Kosinusähnlichkeit unter Umständen unterschiedlich stark selbst bei gleichem Vorkommen eines bestimmten Lexems in zwei Vektoren. Konsequenterweise können Vektoren ihren Platz unter den k nächsten Nachbarn verlieren und durch Vektoren ersetzt werden, die andere Klassenzugehörigkeiten in die Abstimmung bringen. Da es sich bei k um einen trainierbaren Parameter handelt, kann sich insgesamt auch die Größe der abstimmenden Gruppe ändern. Bei Implementierungen, die mit Gewichtungen des Votums nach räumlicher Entfernung arbeiten (Gleichung 2.43) kann zwar die Gruppengröße oder die Gruppenzugehörigkeit eines Vektors erhalten bleiben, eine angepasste Gewichtung nach veränderter Kosinusähnlichkeit jedoch das Abstimmungsergebnis beeinflussen.

Beim Rocchio-Klassifikator hingegen handelt es sich zwar prinzipiell um einen linearen Klassifikator (siehe Unterabschnitt 2.4.1.4). Er basiert jedoch auf der Erzeugung von Durchschnittsvektoren der Klassen im Trainingskorpus und ist somit ebenfalls von den beschriebenen Effekten in diesen Vektoren beeinflusst. Die Bildung dieser Vektoren erfolgt, hier wiederholt aus Gleichung 2.36, in der Form

$$\vec{\mu}(c) = \frac{1}{\mid D_c \mid} \sum_{d \in D_c} \vec{v}(d) \tag{4.27}$$

Handelt es sich bei jeder in den Schwerpunktvektor einbezogenen Eingabe $\vec{v}$ um einen fehlerbehaften TF-IDF-Vektor, befindet sich der resultierende Zentroid $\vec{\mu}$ an einer verschobenen Position im Vektorraum. Auf dieser Basis wird eine ebenfalls verschobene Hyperebene (Gleichungen 2.37-2.42) als Entscheidungsgrenze konstruiert. Diesem Effekt kann entgegengenhalten werden, dass ein unbekanntes, zu klassifizierendes Dokument ähnlichen Fehlertermen unterliegt wie die gemittelten Trainingsbeispiele. Hierbei können sich statistische Artefakte auf der Ebene eines einzelnen Dokuments, etwa in Form von seltenen Homographen zu den eigentlichen Merkmalen, auswirken.

4.6.4.3. Künstliche neuronale Netze und Embeddingvektoren

Wie in Abschnit 2.4.4 dargelegt, werten die einzelnen Einheiten eines künstlichen neuronalen Netzes eine Menge von Eingaben lokal durch eine Aktivierungsfunktion aus. Da es sich dabei ausweislich Gleichung 2.44 um gewichtete Summanden

$$o_j = \sum_i w_{ij} x_i \tag{4.28}$$

handelt, gilt für die lokale Entscheidung analog zur linearen Klassifikation allgemein: Fehlerterme in Merkmalswerten können prinzipiell durch angepasste, erlernte Gewichtungen kompensiert werden. Überdies gilt diese Beeinflussung nur für die erste auf den Eingabelayer folgende verdeckte Schicht, und nur, falls ein frequenzbasiertes Maß wie TF-IDF

als Eingabewert gewählt wird. Da in der Praxis neuronale Netze in der Sprachverarbeitung Embeddingvektoren als Eingabe verwenden, hängt das Maß, in dem Flexion und Homographie in der Eingabeschicht Einfluss ausüben können, von der Repräsentation dieser Phänomene in den vortrainierten Vektoren ab.

Dem Entwicklungsziel eines Embeddings entsprechend, sollen die Vektoren Wortklasse und spezifische Kategorien wie Numerus, Tempus oder Person einer Wortform implizit als latente Variablen enthalten. Ist dieses Ziel erfüllt, sollte es einem neuronalen Klassifikator im Umkehrschluss möglich sein, die entsprechenden Wertebereiche passend zu gewichten und auf diesem Wege eine indirekte Lemmatisierung wortformenübergreifend vorzunehmen. In den Vektoren von verschiedenen Wortformen eines Lexems, etwa *Baum* und *Bäume*, könnte beispielsweise der Numerus „herausgewichtet“ werden. Übrig bliebe wünschenswerterweise der den semantischen Gehalt der Wortform kodierende Teil des Vektors. Kann dieser semantische Kern im Vektor identifiziert werden, können nicht nur im Korpus auftretende flektierte Formen als zusammengehörig, sondern sogar bisher unbekannte Vektoren (etwa *Bäumen*) als zum merkmalsrelevanten Lexem gehörig erkannt werden (sofern zu ihnen ein Embeddingvektor vorliegt).

4.6.5. Zusammenfassung

Dieses Unterkapitel formulierte die drei Ebenen der Beeinflussung eines konventionellen Klassifikationssystems durch Flexion und Homographie: Merkmalsauswahl, Konvertierung von Zeichenketten zu numerischen Merkmalen zur Bildung von Dokumentvektoren und Klassifikationsfunktion. Es konnte gezeigt werden, dass

– nicht generell davon ausgegangen werden kann, dass aufsummierte Vorkommen von Lexemen in jedem Fall höhere χ^2-Werte erhalten als jede der zugrundeliegenden flektierten Formen,

– TF-IDF-Werte von Lexemen in Abhängigkeit von ihren absoluten Häufigkeiten unterschiedlich stark von Lemmatisierung beeinflusst werden,

– Vergleiche insbesondere hochdimensionaler Vektoren in Form der Kosinusähnlichkeit durch eine Lemmatisierung unvorhersehbaren Veränderungen auf Basis von Dimensionalität, absoluten Vorkommen und interner Verteilung zwischen flektierten Formen unterliegen

– verschiedene Typen von Klassifikatoren potenziell unterschiedlich robust auf Flexion und Homographie reagieren: Klassifikatoren, die unmittelbar mit Vektorvergleichen arbeiten, stoßen möglicherweise auf größere Schwierigkeiten als die größere Gruppe der linearen Klassifikatoren, die durch die zu erlernenden Gewichtungen die Verzerrungen der fehlerbehafteten Werte eventuell kompensieren können.

Diese Feststellungen zeigen, dass auf jeder Ebene der Klassifikationsarchitektur Einflüsse der Flexion und der von ihr beeinflussten Homographie im Korpus zu erwarten sind. Sie sind allerdings nur empirisch an einem Beispielkorpus quantifizierbar, da sie sich einer unmittelbar analytischen Lösung entziehen.

Neuronale Netze ohne Vorschaltung einer expliziten Merkmalsauswahl profitieren mutmaßlich vom Vorhandensein von Informationen zur Flexion in dem Embeddingvektor, der die Wortform im Dokument ersetzt: Ist für das Klassifikationsziel nur der semantische Gehalt einer Wortform relevant, kann möglicherweise durch Heruntergewichten der morphologiebezogenen Dimensionen eine implizite Lemmatisierung betrieben werden, die sogar unbekannte flektierte Formen im Zieldokument als zum Lexem gehörig identifizierbar werden lässt. Unklar ist an dieser Stelle noch der Einfluss von Homographie auf den Klasifikator bei Verwendung eines Unigramm-Embeddings wie FastText: Da jede Wortform durch genau einen Vektor repräsentiert wird, fallen homographe Formen verschiedener Lexeme in einem Vektor zusammen. Unter Umständen schlägt sich

die Existenz dieser Homographen allerdings ebenfalls im Vektor nieder und kann bei der Gewichtung des homographiebelasteten Merkmals berücksichtigt werden.

4.7. Schlussbemerkungen und Hypothesen

Die Unterkapitel 4.1 bis 4.5 erläuterten die Erscheinungen der Flexion der deutschen Sprache und begründeten, weshalb Homographie und trennbare Verbpartikeln im Zusammenhang mit Textklassifikation als von Flexion beeinflusst untersucht werden sollen. Teil jedes Unterabschnittes war eine Korpusanalyse, die zeigte, dass die Werte der Kategorien der Wortklassen in der Regel alles andere als gleichverteilt sind. Auch wenn ein Teil dieser Ungleichverteilungen durch Synkretismen bei der Wortformenbildung entschärft wird, stellt sich die Frage, wie ein Klassifikator mit dem Umstand umgeht, dass ein bedeutungstragendes Lexem in Trainings- und Testkorpus potentiell in einer Reihe ungleich verteilter Wortformen auftritt, deren gemeinsamer semantischer Gehalt ohne weitere Informationen nicht bekannt sein kann. Die tatsächliche Realisierung eines potentiellen Formenreichtums im Korpus scheint entscheidend für die Entwicklung des Merkmalsraums, die Formulierung der Texte als Vektoren und das Modellieren der inhaltlichen Kategorien durch den Klassifikator zu sein. In diesem Kapitel wurde somit begründet, welche Einflüsse die Flexion der deutschen Sprache auf den Klassifikationsprozess nehmen kann und dass diese Einflüsse nicht allgemein formal quantifiziert werden können, sondern stets korpus- und szenariospezifisch realisiert werden. Daher werden im folgenden Kapitel in dem aus der Treebank TübaDZ erstellten, in Kapitel 3 beschriebenen Klassifikationsszenario alternativ zu einer ausschließlich analytischen Besprechung die Flexionseffekte in der Klassifikation exemplarisch empirisch untersucht. Erst diese empirische Evaluation kann auch die Korpusgröße als zentralen, bisher nicht besprochenen Einfluss auf den Klassfikationserfolg unter diesem Gesichtspunkt sichtbar machen:

– Gerade starke Ungleichverteilungen einer Flexionskategorie könnten bei kleinen Korpusgrößen zu Artefaktbildung führen: Die Wahrscheinlichkeit, dass ein an sich bedeutungsrelevantes Lexem in einer im Korpus seltenen Form, etwa 2. Person Singular Konjunktiv Präsens Aktiv, auftritt, und statistisch zutreffend interpretiert werden kann, steigt mit zunehmender Korpusgröße. Tritt sie als Artefakt bei einer geringen Korpusgröße auf, kann sie etwa in Form eines χ^2-Wertes fehlinterpretiert werden.

– Größere Merkmalsräume bieten mehr Platz für flektierte Formen; lemmatisierte Texte bilden möglicherweise „dichtere“ Dokumentvektoren. Die Aussagekraft sowohl von mit flektierten Formen besetzten längeren Vektoren als auch kondensierten Lemmata-Vektoren könnte mit zunehmender Korpusgröße statistisch besser abgesichert sein: Flektierte Formen werden umfänglicher in ihrer korrekten Verteilung erfasst, lemmatisierte Vektoren können mehr verdichtete Informationen aufnehmen. Größere Merkmalsräume, realisiert in höherdimensionalen Vektoren, könnten jedoch eine Tendenz zur Verrauschung, das heißt Auffüllung mit irrelevanten Merkmalen, aufweisen, wenn sie nicht aus angemessen großen Trainingskorpora gebildet werden. Im Extremfall kann die Dimension der Dokumentvektoren die Anzahl der Types im Trainingskorpus erreichen und somit Overfitting geradezu erzwingen. Vorteilhaft erscheint daher ein empirisch zu bestimmendes Gleichgewicht zwischen der Vektordimension und der zugrundegelegten Trainings- und somit Informationsmenge.

– Zu prüfen ist, ob in niedrigdimensionalen und/oder auf geringen Korpusgrößen trainierten Klassifikationsmodellen lemmatisierte oder auf andere Arten dimensionsreduzierte Merkmalsräume größeren Klassifikationserfolg ermöglichen und sich dieser Effekt mit zunehmender Merkmalsraumgröße und steigender Trainingsgrundlage abschwächt.

– Auf dem Weg zu einem somit wünschenswert erscheinenden optimalen Gleichgewicht zwischen Merkmalsraumdichte und -größe sind weitere Alternativen zur Lemmatisierung zu prüfen. Denkbar erscheinen etwa die Kennzeichnung der Wortklassen als solche, die

Teillemmatisierung nach Wortklassen und die gezielte Elimination von Homographie sowie die Kombination aus Wortform und Lemma in einem gemeinsamen Merkmal.

– Interessant erscheint die Frage, mit welcher Art Informationen durch Lemmatisierung „freiwerdende“ Stellen in Vektoren bei gleichbleibender Merkmalsraum- und Trainingskorpusgröße aufgefüllt werden: Neben der naheliegenden Vermutung, größere oder lemmatisierte Merkmalsräume ermöglichten eine Verschiebung des Wortklassenanteils zugunsten stärker flektierender Lexeme, soll alternativ in Betracht gezogen werden, dass zusätzliche Plätze im Merkmalsraum durch seltenere, aber spezialisiertere Merkmale, gefüllt werden. In diesem Fall wäre die Reduktion der Flexion oder das Zurverfügungstellen zusätzlicher Plätze von einer semantischen Expansion gefolgt.

5. Experimentelle Untersuchungen

Dieses Kapitel dokumentiert die Experimente, die zur Überprüfung der im vorhergehenden Kapitel getroffenen Annahmen durchgeführt wurden. Der Aufteilung aus Kapitel 2 folgend wird hierbei unterschieden zwischen Experimenten zur Merkmalsauswahl in Unterkapitel 5.1 und Klassifikationsexperimenten auf dem TüBaDZ-Korpus in den Unterkapiteln 5.2 (konventionelle Klassifikatoren) und 5.3 (künstliches neuronales Netz).

5.1. Untersuchungen zur Merkmalsauswahl

Dieses Unterkapitel analysiert in drei Abschnitten die Entwicklung nach dem χ^2-Verfahren zusammengestellter Merkmalsräume. Im ersten Abschnitt erfolgt eine Wortklassenanalyse unter Bezugnahme auf die in Kapitel 4 besprochenen quantitativen Zusammenhänge im Korpus. Ergänzt wird diese Analyse im darauf folgenden Abschnitt um eine Darstellung des Lernverlaufs der flektierten Formen pro Lexem für die offenen Wortklassen. Im dritten Abschnitt wird über eine Analyse zur Entwicklung der internen Clusterdichte nach Kosinusähnlichkeit der zum Merkmalsraum gehörenden FastText-Vektoren eine Perspektive auf die Dynamik in einem Embedding-Vektorraum aufgezeigt.

5.1.1. Wortklassenanalyse

Kapitel 4 dokumentiert das Potenzial der Flexionsphänomene der Deklination und Konjugation sowie flexionsbedingter Homographie, den Prozess der automatischen Klassifikation zu beeinflussen. Die Formenvielfalt der deutschen Verben wurde als potenzielles Problem für den Lern- und Klassifikationsprozess identifiziert, während der größere Anteil der Substantive an der Menge der Types im Untersuchungskorpus auch bei geringerer Flexionsbandbreite ebenfalls Konfliktpotenzial vermuten lässt. Der als Teilbereich der Konjugation betrachtete Komplex der abtrennbaren Verbpartikeln und die flexionsbedingte Homographie wiederum teilen die Eigenschaft, dass bei einer geringeren Menge betroffener Types eine auf Textebene an Ubiquität grenzende weite Verbreitung dieser Phänomene zu konstatieren ist. Bereits diese Betrachtung zeigt auf, dass quantitative empirische Stärke einer Wortklasse im Korpus nicht unbedingt vorauseilend als äquivalent zu großem potenziellem Einfluss auf den Klassifikationsprozess verstanden werden soll.

Sowohl der Trainings- als auch der Operativbetrieb eines Klassifikators vollziehen sich lediglich indirekt auf dem Korpus: Ein Klassifikator trainiert die Interpretation eines explizit durch eine Vorverarbeitung oder implizit in vorgelagerten Layern eines neuronalen Netzes gebildeten Merkmalsraumes, der im Wesentlichen aus der Korrelation zwischen Wortformen (einer beliebigen Verarbeitungsstufe) und Zielklassen extrahiert wurde. Im Produktivbetrieb werden in jedem Fall nur die im Training als relevant identifizierten Merkmale aus dem zu klassifizierenden Text extrahiert und berücksichtigt. Dies ist unabhängig davon, ob es durch aktives Verwerfen in einem konventionellen oder durch Heruntergewichten in einem neuronalen Klassifikationsmodell erfolgt.

Die wortklassenspezifische Formenvielfalt und quantitative Verteilung der einzelnen Wortklassen und Homographen im Korpus wird somit nicht notwendigerweise unmittelbar proportional in den Merkmalsraum übertragen. Die Auswirkungen der

identifizierten Flexionsphänomene sind aus den bisherigen Beobachtungen daher weder quantitativ noch qualitativ abschätzbar. Eine Aussage zur Relevanz einzelner Flexionsaspekte ist daher nicht auf Korpus-, sondern vielmehr auf Merkmalsraumebene zu treffen. Tabelle 5.1 und Abbildung 5.1 zeigen den Anteil der verschiedenen Wortklassen sowie der Homographen an unlemmatisierten Merkmalsräumen exemplarisch für die Kategorie „Sport“. Die Darstellung erfolgt in den Größen 10, 20, 50, 100, 250, 500, 1.000, 2.000, 5.000, 10.000, 20.000 und 50.000 Merkmale. Die Merkmale wurden hierbei nach dem χ^2-Verfahren ausgewählt; die zunehmende Schrittweite bei der Vergrößerung des Merkmalsraums wurde unter Rechenkapazitätsbeschränkungen gewählt. Die Auszählung nimmt bewusst die Perspektive eines unvollständig informierten Merkmalsauswahlalgorithmus, der über keine morphosyntaktischen Zusatzinformationen verfügt, ein, so dass die Kurven der offenen Wortklassen Mehrfachzählungen einzelner Types zwischen diesen Klassen enthalten. Aus demselben Grund sind sämtliche in der Kurve „Homographen“ berücksichtigten Types in den Kurven der offenen Wortklassen bereits enthalten, so dass diese Kurve separat zu betrachten ist.

Merkmale	Substantive	Verben	Adjektive	Andere	Homographen
10	1,00	0,00	0,00	0,00	0,00
20	1,00	0,05	0,00	0,00	0,05
50	0,96	0,08	0,06	0,00	0,10
100	0,95	0,09	0,05	0,00	0,09
250	0,93	0,08	0,04	0,00	0,06
500	0,91	0,06	0,06	0,03	0,05
1.000	0,87	0,06	0,07	0,04	0,04
2.000	0,83	0,07	0,09	0,04	0,04
5.000	0,73	0,10	0,13	0,06	0,03
10.000	0,66	0,16	0,17	0,06	0,06
20.000	0,65	0,20	0,19	0,05	0,10
50.000	0,65	0,18	0,19	0,05	0,08

Tabelle 5.1.: Anteil der Wortklassen im Merkmalsraum, Kategorie „Sport“, unlemmatisiert

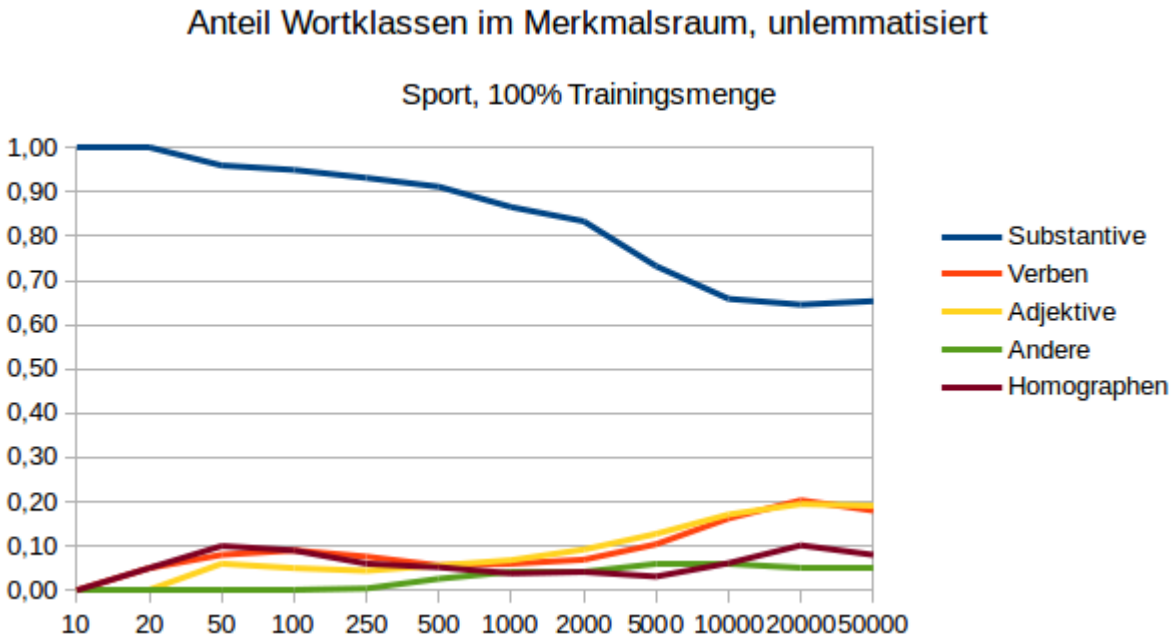


Abbildung 5.1.: Anteil der Wortklassen im Merkmalsraum, Sport, unlemmatisiert

Die Darstellungen zeigen repräsentativ auch für die übrigen Kategorien einen deutlich überproportionalen Anteil der Substantive an den ausgewählten Merkmalen: Ihr Anteil von zunächst 100% liegt bei bis zu 500 Merkmalen bei über 90%, bevor er zugunsten der übrigen Klassen mit zunehmender Schrittweite stetig in Richtung des eigentlichen Korpusanteil von gut 54% zu sinken beginnt. Adjektive und Verben profitieren von großen und sehr großen Merkmalsraumgrößen mit stark kongruentem Verlauf. Der Anteil der Homographen steigt ab 5.000 Merkmalen nach einer zwischenzeitlichen Verringerung wieder an und entwickelt sich auffallend proportional zur Kurve der Adjektive und Verben; der Anteil sonstiger Merkmale steigt nur moderat und bleibt insgesamt gering. Diese Entwicklungen erscheinen im Einklang mit der Vermutung aus den Unterkapiteln 4.6 und 4.7, dass die morphologisch produktiveren Klassen der Verben und Adjektive von größeren Merkmalsräumen profitieren: Seltener auftretende flektierte Formen besetzen Plätze auf längeren Merkmalslisten unterhalb der von Substantiven dominierten Regionen, die

durch geringeren Formenreichtum höhere χ^2-Werte erreichen (siehe Abschnitt 4.6.1 und Gleichung 4.2). Der parallele Verlauf der Kurve der Homographen erinnert an die starke Beteiligung typischer Flexionssuffixe wie *-en* (siehe Tabelle 4.36 in Unterkapitel 4.5).

Tabelle 5.2 und Abbildung 5.2 zeigen eine deutlich abweichende Entwicklung der Merkmalsräume im Fall der Lemmatisierung des Korpus:

Merkmale	**Substantive**	**Verben**	**Adjektive**	**Andere**	**Homographen**
10	0,90	0,00	0,00	0,10	0,00
20	0,90	0,00	0,00	0,10	0,00
50	0,82	0,02	0,02	0,16	0,02
100	0,70	0,02	0,06	0,26	0,04
250	0,62	0,03	0,03	0,34	0,04
500	0,54	0,03	0,03	0,42	0,02
1.000	0,50	0,03	0,05	0,44	0,02
2.000	0,49	0,03	0,07	0,42	0,02
5.000	0,50	0,05	0,11	0,34	0,02
10.000	0,51	0,11	0,17	0,24	0,05
20.000	0,53	0,12	0,19	0,20	0,05
50.000	0,55	0,10	0,17	0,21	0,03

Tabelle 5.2.: Anteil der Wortklassen im Merkmalsraum, Kategorie „Sport“, lemmatisiert

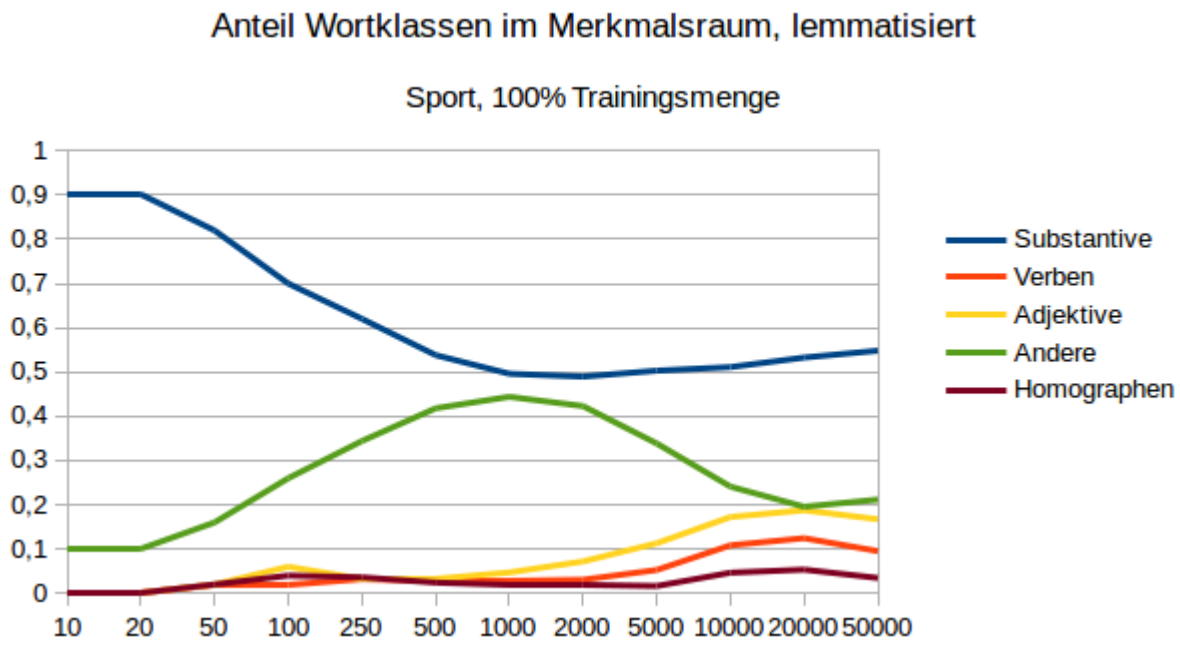


Abbildung 5.2.: Anteil der Wortklassen im Merkmalsraum, Sport, lemmatisiert

Der Anteil der Substantive startet nun niedriger und fällt schneller ab, streckenweise sogar unter ihren Anteil am Gesamtkorpus. Überraschenderweise und entgegen der Annahmen aus Unterkapitel 4.6 profitieren von dieser Verringerung jedoch keineswegs die Verben und Adjektive, deren Aufstieg zwar nach ähnlichem Muster, allerdings mit deutlich verringertem relativem Anstieg parallel einsetzt und nur für die Adjektive zu einem ähnlich hohen Schlusswert führt. Die freiwerdenden Plätze im Merkmalsraum werden vielmehr zum größten Teil von Merkmalen der Kategorie „Andere“ besetzt, die zeitweise nahezu Parität mit den Substantiven erreichen und mit diesen zusammen 94% der Merkmale bei 1.000 Merkmalen stellen. Auch im größten Merkmalsraum hat sich ihr Anteil mit 21% gegenüber der unlemmatisierten Korpusversion mehr als vervierfacht. Diese unerwartete Entwicklung erklärt sich bei einem Blick auf die lemmatisierte Merkmalsliste, die eine große Zahl von Eigennamen unter den sonstigen Merkmalen offenbart, in der Kategorie „Sport“ etwa *devils, lübeck, hsv, steenken, crocodies, brighton, lions, thw, nürn-*

berg. Neben diesen Eigennamen von Personen, Mannschaften und Orten finden sich unter ihnen zusätzlich beispielsweise Zahlen als separierte Merkmale von Mannschaftsnamen. Tatsächlich scheint statt einer Verringerung des Wettbewerbsnachteils der stärker flektierenden Wortklassen Verben und Adjektive eine Verdrängung gewöhnlicher Substantive zugunsten informativerer Eigennamen stattzufinden. Diese treten zwar möglicherweise seltener auf, weisen dann aber eine stärkere Korrelation zur Klasse auf.

Substantive sind in unlemmatisierten Merkmalsräumen überrepräsentiert; dies trifft in größeren Merkmalsräumen auch auf die Adjektive zu. Beide Klassen stellen zusammen den weit überwiegenden Anteil der Merkmale in Größenordnungen von 80-100% über alle Merkmalsraumgrößen hinweg. Somit erscheint eine Vormerkung der Deklination als womöglich dominantem Phänomen über weite Strecken der Merkmalsraumentwicklung und der verbundenen Klassifikationserfolge angebracht. In dem Maße, in dem diese Vermutung bestätigt werden kann, sollten deklinationsvermeidende oder -kompensierende Modifikationen am Korpus zur Steigerung des Klassifikationserfolges führen. Mit zunehmendem Eintrag von konjugierten Merkmalen in größeren unlemmatisierten Merkmalsräumen wäre deren wachsender Einfluss auf Kosten vor allem der Substantive und seine Kompensation von zunehmendem Interesse. Zuletzt ist aus dieser Perspektive den Homographen trotz ihrer im Schnitt mehr als fünffachen Überrepräsentiertheit unter den Types und ihrer vorstehend erhobenen Ubiquität auf Textebene bei Werten um etwa 8% in den größeren unlemmatisierten Merkmalsräumen ein vergleichsweise geringer Einfluss auf den Klassifikationsprozess vorauszusagen. Überdies verringert sich von diesem niedrigen Niveau aus nach einer Lemmatisierung wie in Unterkapitel 4.5 für die Korpusebene beschrieben ihr Anteil erwartungsgemäß im Schnitt über alle Merkmalsgraumgrößen zusätzlich noch um mehr als die Hälfte.

5.1.2. Wortformen pro Lexem

Die in Unterkapitel 4.6 getroffenen Überlegungen zur Beeinflussung des Merkmalsraums beinhalteten die Vermutung, das dem Klassifikator vorgelagerte Merkmalsauswahlverfahren könne mit zunehmender Menge des Trainingsmaterials die statistischen Zusammenhänge zwischen einem bedeutungstragenden Lexem und seinen verschiedenen flektierten Erscheinungsformen im Dokument einerseits sowie der jeweiligen Kategorie andererseits mit wachsender Zuverlässigkeit modellieren. Ein solcher Lernvorgang ließe sich offensichtlich quantitativ in Form der Quote von Wortformen und zugrundeliegenden Lexemen im Merkmalsraum erfassen: Mit einer wachsenden Menge verfügbarer Trainingsdokumente sollte also der Quotient von Wortformen zu Lexemen steigen. Die Abbildungen 5.3 und 5.4 zeigen exemplarisch die Entwicklung dieses Quotienten stellvertretend anhand der Kategorien „Sport“und „Politik“:

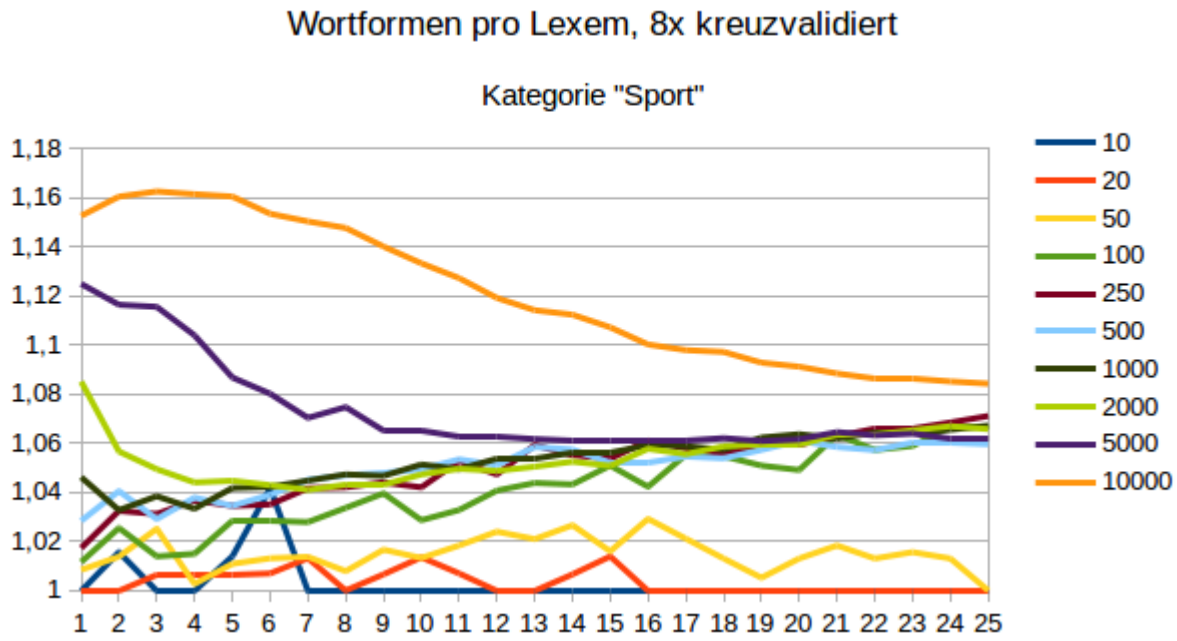


Abbildung 5.3.: Entwicklung Wortformen pro Lexem im Merkmalsraum, Kategorie „Sport“, 25 Datenpunkte zu je 4% Trainingsmengengröße

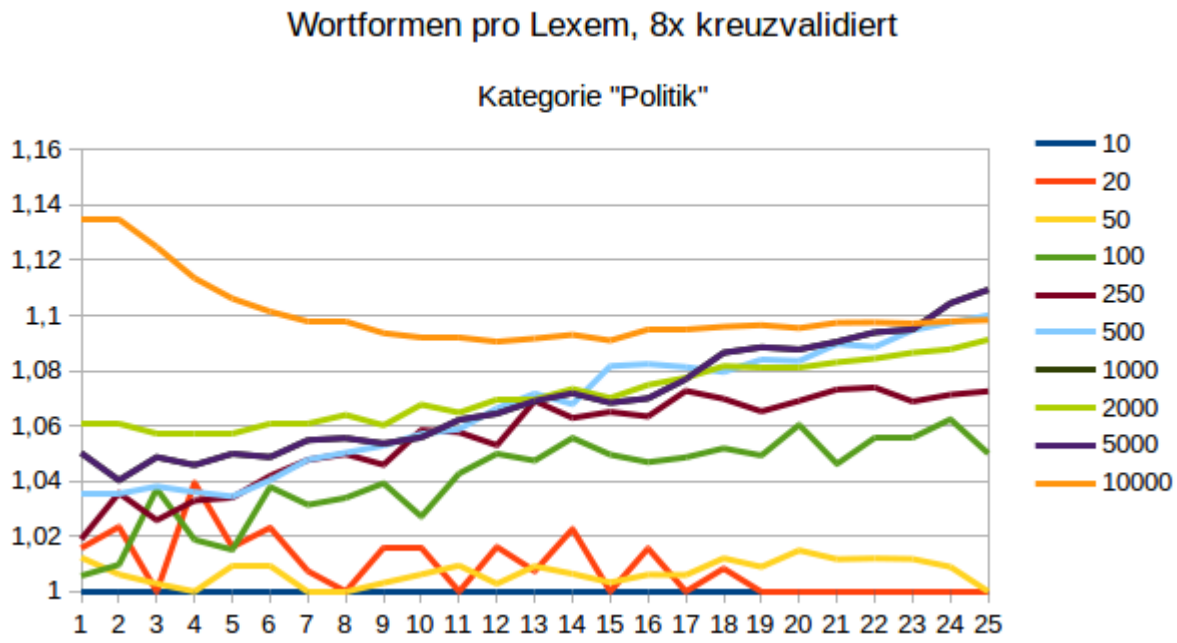


Abbildung 5.4.: Entwicklung Wortformen pro Lexem im Merkmalsraum, Kategorie „Politik“, 25 Datenpunkte zu je 4% Trainingsmengengröße

Die Entwicklung des Verhältnisses von Lexemen zu Wortformen verläuft kategorienübergreifend in gewissem Maße kongruent: Die abnehmende Fluktuation mit zunehmender Merkmalsraumgröße beruht dabei unter anderem auf der Elimination überproportionalen Einflusses als Artefakte auftretender Homographen in den kleineren Merkmalsräumen. Relevanter für den Untersuchungszweck erscheint der ähnliche Verlauf der Quotienten als Funktion der Größe der Trainingsmenge: Sämtliche Kategorien finden sich bei Verwendung des gesamten Trainingskorpus in einem Korridor von etwa 1,05-1,13 Wortformen pro Lexem in den großen Merkmalsräumen ein (mit Ausnahme der Kategorien „Konflikte Ausland“, die teilweise leicht erhöhte Werte erreicht). Die Varianz zwischen

den Kategorien ist gegenüber dem vorhergehenden Experiment deutlich erhöht, und nicht in jedem Fall lässt sich am Ende der berechenbaren Merkmalsraumgröße sicher erwarten, dass die Entwicklung des Quotienten bereits zu einem Abschluss gekommen ist. Auch ohne wortklassenspezifische Auflösung durch Abgleich mit den in Kapitel 4 erhobenen Kennzahlen lässt sich jedoch konstatieren, dass diese Zahlen kategorienübergreifend hinter dem korpusweiten Verhältnis zwischen Lexemen und Wortformen von gut 1:1,3 signifikant zurückbleiben. Gleichwohl korrespondiert gerade die Beobachtung, dass selbst nach Berücksichtigung sämtlicher verfügbarer Trainingsdokumente nicht alle verfügbaren Formen klassifikationsrelevanter Lexeme in den Merkmalsraum übernommen werden, mit der in Kapitel 4 konstatierten deutlich ungleichen Verteilung der durch Flexion gebildeten Types, etwa unter den Gesichtspunkten Kasus, Numerus und Tempus. Bemerkenswert erscheint überdies, dass sich der relative Zuwachs beim Quotienten Wortformen/Lexeme mit zunehmender Merkmalsraumgröße verlangsamt: Die zusätzlichen Stellen im Merkmalsraum scheinen sowohl nach absoluten als auch relativen Werten statt durch morphologische Zusatzinformationen eher durch semantische Ergänzungen gefüllt zu werden. Der Umstand, dass in sämtlichen, auch großen, Merkmalsräumen der Quotient von Wortformen und Lexemen, wenn auch mit steigender Tendenz, deutlich hinter dem Verhältnis im Korpus insgesamt zurückbleibt, lässt vermuten, dass parallel zu dieser lediglich moderaten morphologischen Diversifizierung semantische Aspekte eine konkurrierende Rolle bei der Gestaltung des Merkmalsraums in Abhängigkeit zur Größe der Trainingsmenge spielen. Eine Annäherung an dieses Verhältnis zwischen semantischen und morphologischen Aspekten als Merkmale verwendeter Wortformen und Lemmata soll im folgenden Abschnitt aus der Perspektive ihrer FastText-Embeddings erfolgen.

5.1.3. Clusterdichte FastText

Statistische Merkmalsauswahlkriterien wie der hier verwendete χ^2-Wert treffen lediglich eine Aussage über die Stärke der Korrelation eines atomaren Symbols und einer abhängigen Variablen. Im Falle der Textkategorisierung ist dies die Korrelation einer Wortform und einer Zielklasse. Eine Aussage über einen wie auch immer definierten semantischen Gehalt oder über morphologische Eigenschaften des Symbols, also des Merkmals, das dem Klassifikator gegebenenfalls übergeben wird, ist hieraus nicht abzuleiten. Im Unterschied zu dieser Vorgehensweise zielt das Embeddingkonzept, wie in Abschnitt 2.4.4 erläutert, darauf ab, einem Symbol einen multidimensionalen reellwertigen Vektor zuzuweisen, der dessen kontextgebundenes Vorkommen in einem Trainingskorpus beschreibt und sowohl semantische als auch morphologische Informationen in Form latenter Variablen enthält. Die verschiedenen Erscheinungsformen eines Lexems sollten also durch Vektoren repräsentiert werden, die einander ähnlicher sind als denen nicht verwandter Lexeme. Ebenso sollten semantisch zueinander in Beziehung stehende Lexeme beziehungsweise deren Umsetzungen in Wortformen eine größere Ähnlichkeit untereinander aufweisen als semantisch fremde Begriffe.

Die Kosinusähnlichkeit wurde als in NLP-Kontexten etablierte Metrik zum Vergleich zweier Vektoren vorgestellt. Weisen allgemein die Embedding-Vektoren zweier semantisch oder morphologisch verwandter Wortformen eine größere Kosinusähnlichkeit auf als die zufällig ausgewählter Paare von Wortformen, so sollten die wichtigsten Merkmale einer Kategorie untereinander eine durchschnittlich größere Kosinusähnlichkeit aufweisen als im Vergleich mit den Merkmalen anderer Kategorien oder dem Korpus ingesamt. Das im Folgenden vorgestellte Experiment vergleicht die Vektoren jeweils sämtlicher Merkmale der acht Kategorien untereinander in den bereits eingeführten Merkmalsraumgrößen. Die Metrik ist hierbei der in SciPy implementierte Kosinusabstand, der dem Wert $1 - Kosinus\text{ähnlichkeit}$ entspricht. Nach diesem wird

jedes Merkmal eines Merkmalsraums mit sämtlichen anderen Mitgliedern desselben Raumes verglichen und auf die Anzahl dieser Vergleiche normalisiert, so dass sich nach $n^2 - n$ Vergleichen ein durchschnittlicher Kosinusabstand von

$$KA = \frac{\sum_{i=1}^{n} 1 - \cos(\vec{u_i}, \vec{v_i})}{n^2 - n} = \frac{\sum_{i=1}^{n} 1 - \frac{\vec{u_i} \cdot \vec{v_i}}{\|\vec{u_i}\| \|\vec{v_i}\|}}{n^2 - n} \tag{5.1}$$

ergibt. In der Normalisierung erfolgt eine Korrektur der Anzahl n^2 der aufgeführten Vergleiche um n, da der Kosinusabstand eines Merkmals mit sich selbst zwar stets 0 beträgt und somit nicht direkt in den durchschnittlichen Abstand einfließt, wohl aber die Anzahl der Vergleiche erhöht; dieser Umstand wirkt sich in den kleineren Merkmalsräumen deutlicher verzerrend aus als in den großen Merkmalsräumen ($10^2 - 10 = 90$ vs. $5000^2 - 5000 = 24,995$ Mio; dies entspricht 10% im Vergleich zu 0,5%). Des Weiteren ist zu beachten, dass in diesem Experiment ein Vergleich nur zwischen den Vektoren jener Merkmale gezogen werden kann, die auch in FastText aufgefunden werden. Bei n handelt es sich mithin nicht um die nominale Merkmalsraumgröße, sondern um die Anzahl gefundener Types. Der Anteil in FastText gefundener Types eines Merkmalsraums schwankt erheblich mit dessen Größe und der Anzahl der vorgestellten Trainingstexte. Die Abbildungen 5.5 und 5.6 zeigen die Entwicklung dieses Anteils exemplarisch für die Kategorie Kategorie „Umwelt“.

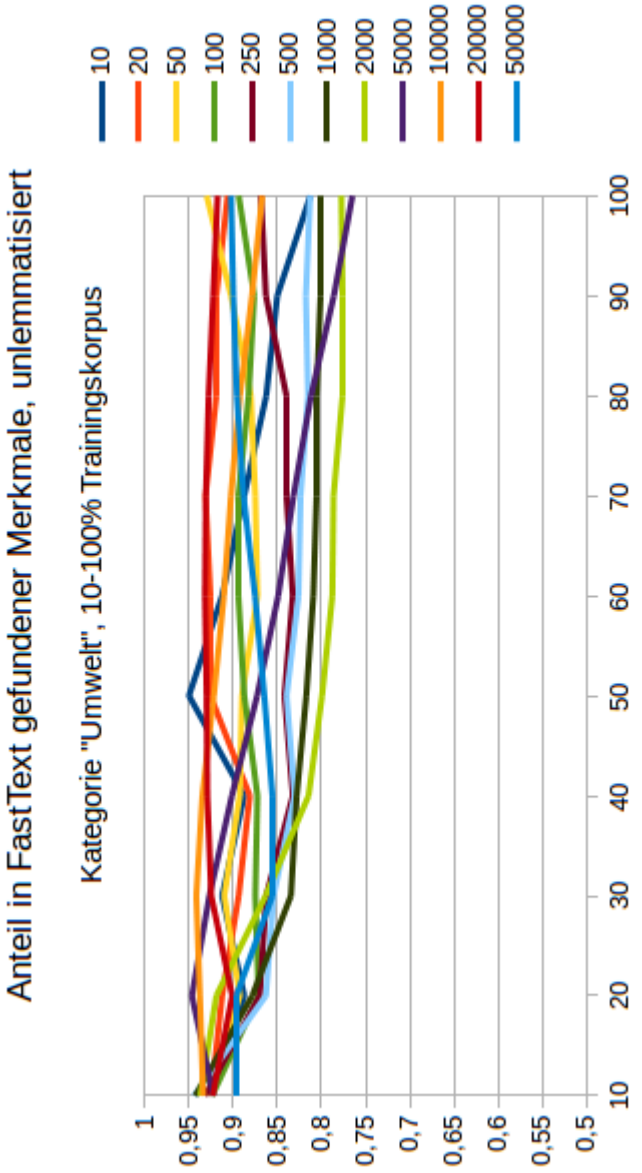


Abbildung 5.5.: Anteil in FastText gefundener Merkmale, Kategorie „Umwelt“, unlemmatisiert

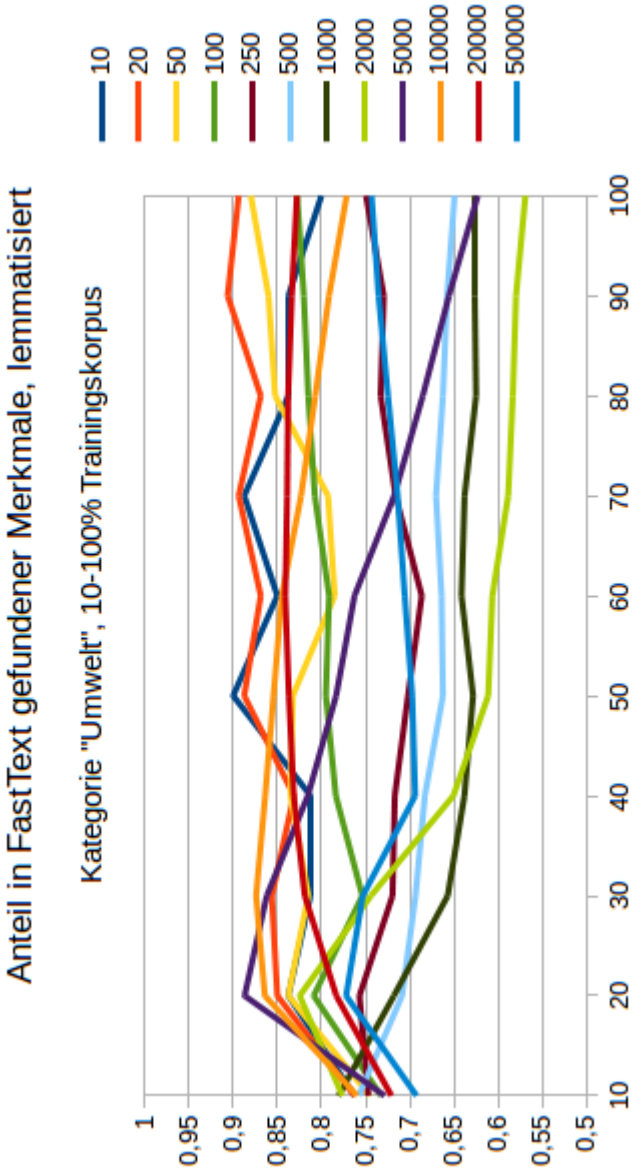


Abbildung 5.6.: Anteil in FastText gefundener Merkmale, Kategorie „Umwelt“, lemmatisiert

Erkennbar sind zwei wesentliche Trends über Merkmalsraumgröße und Trainingsmenge: Zum einen sinkt im Allgemeinen der Anteil der gefundenen Terme mit zunehmender Merkmalsraumgröße. Dieser Trend beruht auf dem simplen Frequenzeffekt, dass größere Merkmalsräume per se mit zunehmend selteneren Termen ergänzt werden. Bei

der zweiten sichtbaren Entwicklung ist zwischen unlemmatisierten und lemmatisierten Merkmalsräumen zu unterscheiden: Während der Anteil der im Embedding gefundenen unlemmatisierten Merkmale im Wesentlichen mit wachsendem Trainingskorpus in jeder Merkmalsraumgröße abnimmt, betrifft dies bei den lemmatisierten Merkmalen vorwiegend die größeren Merkmalsräume. Die Entwicklung bei den lemmatisierten Merkmalen kann dem Umstand geschuldet sein, dass das χ^2-Filterkriterium mit zunehmender Merkmalsraumgröße Terme mit hoher Precision auch bei geringem Recall, hier proportional zu geringer Frequenz, in den Merkmalsraum befördert. Diese Terme weisen zwar eine hohe Spezifizität für die Kategorie auf, erscheinen aber nur in wenigen Dokumenten und/oder sehr selten insgesamt. Seltene Terme sind aber offensichtlich mit geringerer Wahrscheinlichkeit im vortrainierten Embedding zu finden. In unlemmatisierten Merkmalsräumen hingegen werden freie Plätze mutmaßlich in einem gewissen Ausmaß erst noch mit flektierten Formen der bedeutungstragenden Lexeme aufgefüllt, die häufiger im Embedding gefunden werden.

Schließlich ist zu bemerken, dass allgemein der Anteil der aufgefundenen Terme in den unlemmatisierten Merkmalslisten in der Mehrheit der Fälle moderat bis deutlich höher ist als in den lemmatisierten: Das Training der Standarddistribution von FastText auf CommonCrawl und Wikipedia scheint nicht notwendigerweise die Zitierform von Lexemen, sondern die im überwiegend schriftsprachlichen, relativ formalen Eingabematerial abweichend verteilten flektierten Formen zu bevorzugen. In mehreren flektierten Formen vorkommende höherfrequente Lexeme werden möglicherweise gegenüber selteneren, nur in der Zitierform auftretenden domänenspezifischen Lexemen des Untersuchungskorpus bevorzugt. Das Substantivkompositum „Zweitligakandidat“ beispielsweise erscheint im Korpus ein einziges Mal, im Nominativ Singular, lässt eine hohe Spezifität für die Kategorie „Sport“ erwarten und ist nicht in FastText enthalten (zum Vergleich: FastText enthält Vektoren zu 78 auf „-kandidat“ endenden und zu 94 auf „-kandidaten“ endenden Wortformen).

Abbildung 5.7 zeigt am Beispiel der Kategorie Kategorie „Wirtschaft“exemplarisch für die kategorienübergreifende Entwicklung den nach vorstehend beschriebenem Verfahren ermittelten durchschnittlichen Kosinusabstand nach Merkmalsraumgröße als Funktion der Größe der Trainingsmenge.

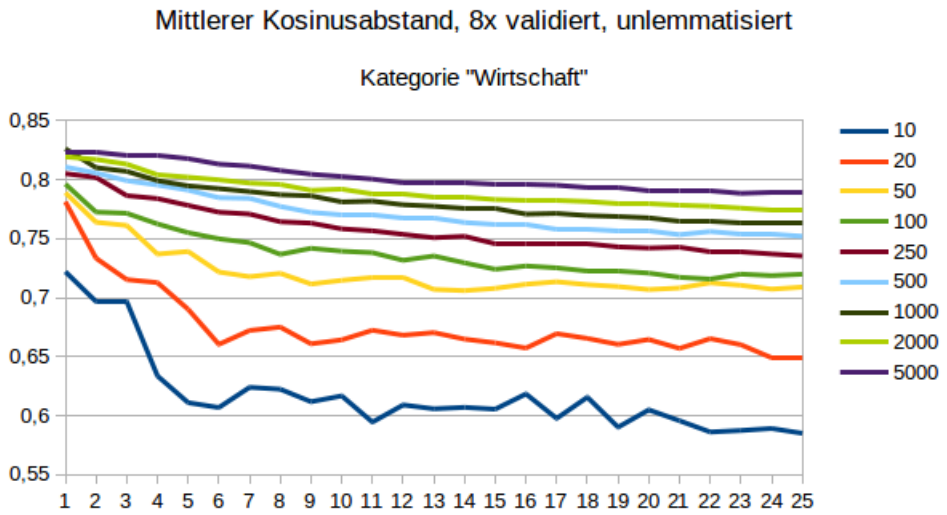


Abbildung 5.7.: Durchschnittlicher Kosinusabstand, unlemmatisiert, Kategorie „Wirtschaft“, 25 Datenpunkte Korpusgröße zu je 4%

Wie bei den beiden vorangegangenen Experimenten ergeben sich deutliche kategorieübergreifende Parallelen mit wesentlichen nachvollziehbaren Trends: Erwartungsgemäß zeigen sich ebenfalls abnehmende Fluktuation in größeren Merkmalsräumen, semantisch-morphologische Diversifikation mit zunehmender Trainingsmenge, größere Diversität in dieser Hinsicht in kleineren Merkmalsräumen und Konvergenztendenzen beim durchschnittlichen Kosinusabstand.

Da, wie bereits erwähnt, die der Abstandsmessung zugrundeliegenden Merkmalsvektoren morphologische wie semantische Informationen beinhalten, kann bei der hier sichtbaren teils deutlichen Verdichtung der Kategorien mit zunehmender Trainingsmenge nicht

ohne Weiteres zwischen diesen beiden Einflüssen unterschieden werden: Die im vorhergehenden Experiment beobachtete moderate morphologische Diversifikation könnte, wie bereits überlegt, von semantischer Expansion begleitet und von dieser nach der aktuellen Metrik teilweise kompensiert werden.

Aufschluss über den Anteil morphologischer Diversifikation von Lexemen kann eine Wiederholung des Experiments auf einem ausschließlich aus Lemmata bestehenden Merkmalsraum geben. Die Ergebnisse dieser Variante des Experiments zeigt beispielhaft anhand der Kategorie Kategorie „Wirtschaft“ Abbildung 5.8.

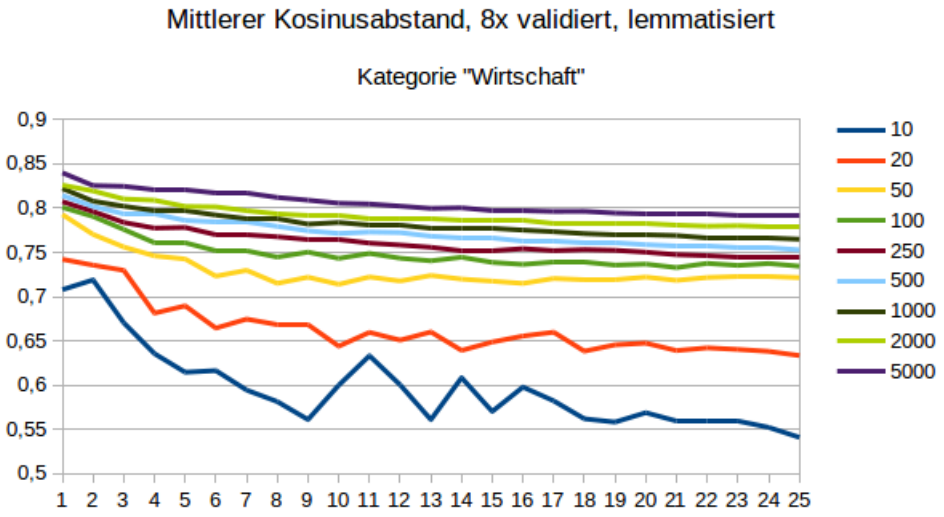


Abbildung 5.8.: Durchschnittlicher Kosinusabstand, lemmatisiert, Kategorie „Wirtschaft“, 25 Datenpunkte Korpusgröße zu je 4%

Die zu den unlemmatisierten Merkmalslisten beschriebenen Tendenzen im Verhältnis zur Trainingsmenge lassen sich ebenfalls kategorienübergreifend auch für die Entwicklung der lemmatisierten Merkmalsräume konstatieren.

Abbildung 5.9 als Metaauswertung der durchschnittlichen Kosinusabstände der jeweils unlemmatisierten und lemmatisierten Merkmalslisten verdeutlicht jedoch, dass auch die

Gegenüberstellung von lemmatisierten und unlemmatisierten Merkmalslisten keinen eindeutigen Aufschluss über ein merkmalsrauminternes Verhältnis von Morphologie und Semantik im Lernprozess gibt:

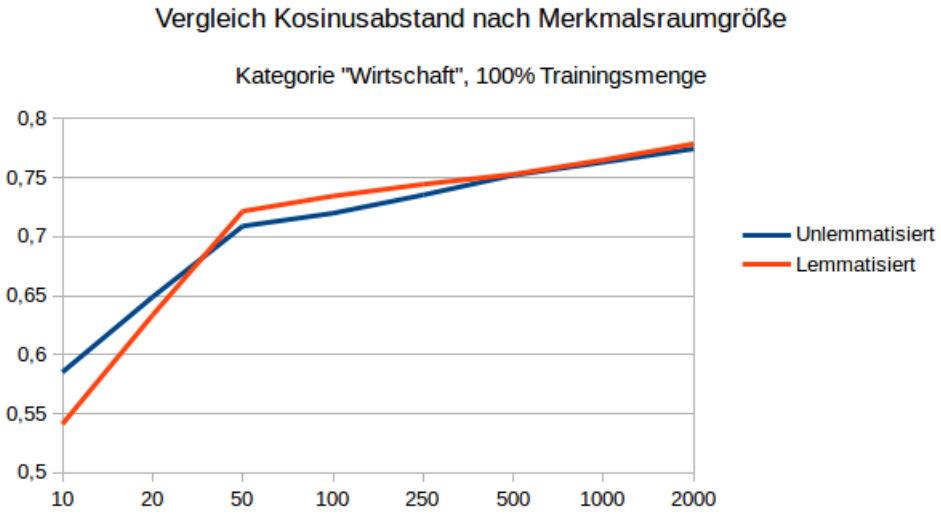


Abbildung 5.9.: Durchschnittlicher Kosinusabstand, unlemmatisiert, Kategorie „Wirtschaft“

Der durchschnittliche Kosinusabstand der lemmatisierten Merkmalslisten liegt bis 20 Merkmale unter dem der unlemmatisierten, zwischen 50 und 500 Merkmalen etwas oberhalb und sodann im Wesentlichen gleichauf. So schwerlich in dieser Kategorie ein eindeutiger Trend zu definieren ist, so wenig repräsentativ ist die Entwicklung dieser Kategorie ohnehin: Anfänglich (fast vollständig) größere Abstände der lemmatisierten Listen weisen lediglich die Kategorien „Konflikte Ausland“, „Kriminalität“ und „Kultur“ auf; die verbleibenden Kategorien „Panorama“, „Politik“, „Sport“ und „Umwelt“ hingegen erzeugen anfänglich dichtere unlemmatisierte Merkmalslisten. Die Kategorien „Kultur“, „Panorama“, „Politik“, „Sport“und „Umwelt“weisen einen hohen Grad an Kongruenz beider Kurven auf, im Fall von „Sport“ sogar nahezu vollständig und von Beginn an.

Diese Ergebnisse zeigen die Schwierigkeiten auf, Voraussagen über lexikalische Entwicklungen und ihre Parallelen in einem manuell nicht analysierbaren hochdimensionalen Vektorraum, namentlich Korrelationen mit morphologischen und semantischen Informationsverdichtungen in einem korpusbasierten Prozess maschinellen Lernens, zu treffen. Gegenstand dieser auf den Einfluss flexionsmorphologischer Prozesse auf den Klassifikationsprozess ausgerichteten Arbeit kann weder eine redaktionelle Analyse der Entwicklung der automatisch erstellten Listen unter diesen interagierenden Gesichtspunkten noch die Skizzierung einer Art Merkmalsraumsemantik lemmatisierter oder unlemmatisierter Texte sein. Dennoch sei eine denkbare Erklärung für die Fluktuationen in den Merkmalsräumen, ihren Abständen untereinander und denen zwischen lemmatisierten und unlemmatisierten Merkmalslisten sowie die große Diversität zwischen den Kategorien hier anhand eines generischen Beispiels zur Kategorie „Sport“ skizziert:

Während das Lexem *Tor* aus Texten zum Thema Fußball dieser Kategorie in den nach χ^2 als aussagekräftig eingestuften flektierten Formen *Tor*, *Tore* und *Toren* drei Plätze in einer unlemmatisierten Merkmalsliste belegen könnte, würde es in einer lemmatabasierten Merkmalsliste lediglich einen einzigen Listenplatz beanspruchen. Die bei in den Experimenten vorausgesetzter fixer Vektorlänge freiwerdenden Plätze der flektierten Formen wären nun ersatzweise durch zwei zusätzliche Lemmata zu besetzen. Nach der Zielsetzung der Embeddingverfahren ist davon auszugehen, dass flektierte Formen eines Lexems im trainierten Vektorraum diesem nicht nur semantisch, sondern mindestens auch durch die Wortklasseninformation sowie indirekt durch die latent markierten wortklassenspezifischen Informationen wie Numerus und Kasus ähnlich sind: Die Beziehung „ist der Plural von“ zwischen *Tore* und *Tor* ist, wie bereits besprochen, indirekt über den Vektorvergleich ermittelbar. Nach diesem Verständnis wären zusätzliche Nachrückerlexeme wie *Schiedsrichter* oder *Bundesliga* dem Lexem *Tor*, hier in seiner Zitierform als Merkmal in der Liste auftretend, per Definition weniger ähnlich, als es dessen flektierte Formen sein können. Dies schließt jedoch nicht aus, dass sich etwa diese Lexeme un-

tereinander oder den weiteren Lexemen der Liste insgesamt ähnlicher sind, als dies bei dem Lexem *Tor* und den bisherigen weiteren Lexemen im bisherigen Merkmalsraum im Durchschnitt der Fall war. Zwar lassen sich nach dem Designprinzip der Embeddingvektoren Ähnlichkeiten zwischen flektierten Formen untereinander sowie zu ihrem Lemma ebenso extrahieren wie semantische Ähnlichkeiten zwischen verschiedenen Lexemen, eine konkrete Metrik oder Quantifizierbarkeit hierfür ist jedoch als semantische Frage nicht mehr Gegenstand dieser Untersuchung.

Die in wechselnden Merkmalsraumgrößen kategorienabhängig alternierend höhere und geringere Dichte unlemmatisierter und lemmatisierter Merkmalslisten nach der Kosinusabstandsmetrik scheint Symptom eines dynamischen Ausgleichsprozesses zwischen morphologischer und semantischer Abdeckung des Inhalts der jeweiligen Kategorie zu sein. Auch wenn semantische Aspekte und detaillierte Korpusanalysen in dieser Hinsicht in dieser Untersuchung zurückstehen sollen, scheint eine Korrelation zwischen diesen Mustern und dem Grad semantischer Unterkategorisierung über Teilcluster (etwa bei den inhaltlich breiteren Kategorien „Panorama“ und „Konflikte Ausland“) naheliegend.

Die vorstehenden Betrachtungen zeigen, dass eine größere Ähnlichkeit lemmatisierter Merkmalslisten als klassifikationsrelevanter Repräsentationen von Texten der jeweiligen Kategorie gegenüber ihren unlemmatisierten Pendants nicht pauschal vorausgesetzt werden kann. Dieses Fazit des embeddingbasierten Merkmalslistenexperiments ergänzt die in Unterkapitel 4.6 getroffenen Feststellung, dass allein aus formalen Gründen nicht vorausgesetzt werden kann, dass vektorisierte lemmatisierte Texte nach TF-IDF einander in jedem Fall ähnlicher sind als unlemmatisierte Texte.

Zu beachten ist, dass die Dichte eines Merkmalsraums im Sinne dieser Metrik keine Voraussage über die Leistungsfähigkeit des Merkmalsraums als Grundlage eines Klassifikators treffen soll: Sowohl konventionelle als auch neuronale Klassifikatoren gewichten die in numerische Werte umgewandelten Merkmale in einer Funktion, mit der die Positi-

on eines Dokuments im Vektorraum auf eine oder mehrere Zielvariablen, die Kategorien, abgebildet werden soll. Die Ähnlichkeit der Merkmale untereinander nach morphologischen oder semantischen Gesichtspunkten ist hierbei mutmaßlich nur ein bedingter Indikator für die Fähigkeit des Klassifikators, eine hinreichende Varianz der Zielkategorie abzubilden, ohne Overfitting zu verursachen. Von den untersuchten Klassifikatoren verwenden lediglich neuronale Netze die Embeddingvektoren direkt und treffen hier auch ihre eigene Auswahl nicht nur der Merkmale an sich, sondern durch Gewichtung einzelner Vektordimensionen zielgerichtet auch die relevanter Aspekte des Merkmals. Diese Vorgehensweise unterscheidet sich signifikant von der dokumentbasierten, auf TF-IDF aufbauenden Arbeitsweise der übrigen untersuchten Klassifikatoren. Die Aussagekraft dieser Merkmalsraumvergleiche soll somit ausschließlich als auf das Verhältnis von Morphologie und Semantik in Merkmalsräumen speziell unter der Interpretation des kosinusbasierten Abstandsmaßes beschränkt verstanden werden; die Fähigkeit der Klassifikatoren, diese Einflüsse zu handhaben und Störungen in der einen oder anderen Dimension zu kompensieren, ist hieraus nicht abzuleiten.

Gleichwohl kann die Feststellung, dass die Lemmatisierung den durchschnittlichen Kosinusabstand der Merkmale einer Kategorie in manchen Fällen signifikant verringert, möglicherweise dahingehend interpretiert werden, dass die Entfernung jeglicher morphologischer Einflüsse bei gleichbleibender Merkmalsraumgröße Raum für die Berücksichtigung auch semantisch speziellerer Merkmale, also für eine inhaltlich umfassendere Modellierung einer Kategorie, bereitstellt. Wie besprochen konkurriert dieses Ziel in der Architektur und Entwicklung eines Klassifikationssystems mit dem Wunsch nach Vermeidung von Overfitting, also der Modellierung zu spezieller Aspekte oder einzelner Datenpunkte einer Kategorie. Welche jeweilige morphologische und semantische Bandbreite und welche Anteile dieser beiden Aspekte an der Gestaltung des Merkmalsraumes für eine vorliegende Problemstellung optimal sind, erscheint vor einer empirischen Untersuchung korpus- und szenariospezifisch: Korpusspezifisch, da jedes Textkorpus sprachlich von Do-

mäne und Register geprägt ist, und szenariospezifisch im Sinne semantischer Dichte der Kategorien inklusive der Aspekte Unterkategorien und Streubreite sowie Ausreißerinstanzen.

Diese Indikationen werden ergänzt von einer in Tabelle 5.3 dargestellten Testreihe, in der die Kosinusabstände zweier flektierter, in FastText abgebildeter Formen von je 500 zufällig aus dem Korpus ausgewählten Adjektiven, Substantiven und Verben (etwa *monumentalen – monumentale* oder *kapitel – kapiteln*), untereinander ermittelt werden.

Wortklasse	**Kosinusabstand**
Adjektive	0,1598
Substantive	0,1940
Verben	0,3140

Tabelle 5.3.: Durchschnittlicher Kosinusabstand flektierter Formen nach Wortklassen, Korpusebene

Der durchschnittliche Kosinusabstand zweier zufälliger flektierter Formen desselben Lexems fällt offensichtlich wortklassenübergreifend deutlich geringer aus als der durchschnittliche Abstand der in gewissem Grad semantisch in Zusammenhang stehenden Lexeme einer Kategorie. Dies lässt darauf schließen, dass morphologische Verwandtschaft in FastText-Vektoren prominent gewichtet ist und die Präsenz weitere flektierter Formen neben bereits vorhandenen Formen eines Lexems den durchschnittlichen Kosinusabstand verringern kann. Dieses Potenzial scheint erheblich, muss aber neben den zwei bereits ermittelten Faktoren eingeordnet werden: Das Potenzial semantischer Verdichtung sowie den wachsenden Einfluss der Verben, die durchschnittlich höhere Abstände aufweisen als die eingangs dominanteren Substantive. Wie weit dieser wiederum durch den geringeren durchschnittlichen Abstand der Adjektive kompensiert wird, bleibt spekulativ. Der Anteil aller drei Effekte an der Entwicklung eines spezifischen Merkmalsraums im vorliegenden oder einem anderen Klassifikationsszenario ist ohne qualitative Analyse nicht quantifizierbar.

5.1.4. Zusammenfassung und Diskussion zu den Untersuchungen zur Merkmalsauswahl

Die Ergebnisse der drei Experimente dieses Unterkapitels ergeben miteinander in Zusammenhang gesetzt ein Bild interagierender Mechanismen bei der Merkmalsraumbildung in einem Nachrichtentextkorpus: Die Wortklassenanalyse in Abschnitt 5.1.1 zeigt, dass die morphologisch weniger produktiven Substantive ihren anfänglichen „Wettbewerbsvorteil“ zugunsten der formenreicheren Wortklassen und der Eigennamen verlieren. Entgegen der Hypothesen in Unterkapitel 4.6 profitieren Verben und Adjektive jedoch von einer Lemmatisierung nicht in jedem Fall in Form einer Belegung zusätzlicher Merkmalsplätze. Jene werden zeitweise ganz überwiegend und generell überproportional häufig von Eigennamen und sonstigen Termen übernommen. Die Wortklassenverteilung im unlemmatisierten Merkmalsraum schlägt sich im Ergebnis des Experiments 5.1.2 zur Anzahl der flektierten Formen pro Lexem nieder: Diese steigt zwar mit zunehmender Trainingsmenge und auch im Verhältnis zur Merkmalsraumgröße, also der Anzahl der für weniger häufige flektierte Formen potenziell zur Verfügung stehenden Listenplätze. Der Anstieg bleibt jedoch deutlich hinter der Quote der Wortformen zu Types und Lexemen im Korpus insgesamt zurück. Zusätzlich flacht der Anstieg der Quote relativ zur Erweiterung der Merkmalsräume stark ab und hält mit dem relativ stark steigenden Anteil der Verben und Adjektive nicht proportional mit. Diese Diskrepanz kann schließlich durch das dritte Experiment zur Clusterdichte, gemessen mit dem durchschnittlichen Kosinusabstand verfügbarer FastText-Vektoren, aufgelöst werden. Das Experiment zeigt, dass der durchschnittliche Kosinusabstand sowohl unlemmatisierter als auch lemmatisierter Merkmalsräume im Trainingsfortschritt in jeder Merkmalsraumgröße abnimmt. Diese Verdichtung kann morphologische wie semantische Gründe haben, deren Verhältnis ohne zusätzliche Informationen nicht quantitativ präzisierbar ist. Aus der beobachteten Entwicklung der Wortklassenverteilung im Zusammenhang mit der Quote der Wortformen pro Lexem kann jedoch sicher geschlossen werden, dass seltenere Lexeme mit

wenigen Formen Beiträge zur semantischen Verdichtung leisten: Plätze, die nicht von morphologischer Diversifikation belegt werden, werden zur semantischen Spezialisierung genutzt.

Auch wenn die drei Experimente einige klar erkennbare, kategorienübergreifend zu beobachtende Trends aufzeigen können, bleibt deren genaue Interaktion in den Merkmalsräumen dieses spezifischen Szenarios unklar und ohne qualitative Analyse nicht weiter auflösbar. Letztere wiederum erscheint für die größeren Merkmalsräume nicht manuell durchführbar und ist nicht zuletzt auch wegen der semantisch-redaktionellen Fragestellungen nicht mehr Gegenstand dieser Untersuchung. Der szenariospezifische Einfluss der Wortklassenverteilung, der Merkmalsraumgröße sowie der Verwendung alternativer Merkmalstypen anstelle flektierter Wortformen auf den Klassifikationserfolg ist Gegenstand der Experimente der folgenden Unterkapitel 5.2 und 5.3. Gleiches gilt für die jeweiligen Lernvorgänge zu diesen Einflussfaktoren im Hinblick auf die zur Verfügung stehende Trainingsmenge.

5.2. Konventionelle Klassifikation

Dieses Kapitel dokumentiert eine Reihe von Experimenten, die zur Überprüfung der in Unterkapitel 4.6 aufgestellten Hypothesen mit verschiedenen Versionen des TübaDZ-Korpus durchgeführt wurden. Abschnitt 5.2.1 stellt neun Modifikationen des TübaDZ-basieren Klassifikationskorpus vor. Abschnitt 5.2.2 dokumentiert die Ergebnisse der sechs konventionellen Klassifikationsverfahren K-nächste-Nachbarn, Logistische Regression, Naive Bayes (Modelle Bernoulli und Multinomial), Rocchio und Support Vector Machine unter Nutzung dieser Korpusvarianten.

5.2.1. Wortformenbasierte Korpusmodifikationen

Basis der experimentellen Untersuchungen zum Einfluss der in Kapitel 4 aufgeführten Flexionsphänomene mit konventionellen Klassifikationsverfahren bilden die im Folgenden vorgestellten Modifikationen des in Kapitel 3 beschriebenen Klassifikationsszenarios. Unterabschnit 5.2.1.1 erläutert die am Originalkorpus vorgenommenen Modifikationen; Unterabschnitt 5.2.1.2 formuliert Hypothesen zu den experimentellen Auswirkungen dieser Modifikationen.

5.2.1.1. Modifikationen des TübaDZ-Nachrichtenkorpus

Die neben dem unlemmatisierten Originalkorpus verwendeten neun modifizierten Varianten können in zwei Gruppen unterteilt werden: Die erste Gruppe besteht aus den Korpusvarianten *Lemmatisiert*, *Wortform plus POS-Tag*, *Lemma plus POS-Tag* und *Wortform plus Lemma*. Diese vier Korpusversionen haben gemeinsam, dass zu ihrer Erstellung keinerlei linguistische Bearbeitungen vorgenommen, sondern lediglich die benötigten Spalten aus dem CONLL-Format extrahiert werden mussten. Die zweite Gruppe umfasst zielgerichtet linguistisch modifizierte Korpusversionen: *Ohne Deklination*, *Ohne Konjugation*, *Ohne abtrennbare Verbpartikeln („Ohne PTKVZ“)*, *Ohne Homographen* sowie *Lemma plus kleines POS-Tag-Set („Lemma+4POS“)*. Die ersten vier Korpusversionen dieser zweiten Gruppe haben die Elimination spezifisch jeweils eines der vier Flexionsphänomene gemeinsam, während die fünfte die aus dem CONLL-Format extrahierten Lemmata um ein verkleinertes Tagset ergänzt. Bei diesem Tagset handelt es sich um Tags für die Wortklassen Substantiv, Verb, Adjektiv oder Sonstige. Dieses reduzierte Tagset kommt auch in der Korpusversion *Ohne Homographen* als Ergänzung der unlemmatisierten Wortformen zum Einsatz. Da entsprechend der im Vorfeld getroffenen einschränkenden Annahme, dass nur Types aus den drei offenen Wortklassen klassifikationsrelevante Informationen bereitstellen,

Homographie nur zwischen diesen Wortklassen untersucht werden soll, genügt zu deren Ausschaltung eine Unterscheidung zwischen diesen Klassen und „Sonstigen“.

Die Ausschaltung der Deklination erfolgt durch die Lemmatisierung ausschließlich der Substantive und Adjektive (Korpusvariante *Ohne Deklination*). Die Ausschaltung der Konjugation erfolgt analog durch die Lemmatisierung der finiten Verbformen, während alle anderen Wortformen unverändert bleiben (Korpusvariante *Ohne Konjugation*). Die Ausschaltung des Phänomens der abtrennbaren Verbpartikeln erfolgt, indem diese an ihr jeweiliges Finitum angefügt werden (Korpusvariante *Ohne PTKVZ*). Tabelle 5.4 stellt diese Modifikationen ergänzt um zwei Beispiele zusammen (abweichendes Beispielverb für Homographen in der letzten Zeile). Durch die Anfügung des vierteiligen Tagsets an die Wortformen eliminiert die Version *Ohne Homographen* wortklassenübergreifende Homographie unter Beibehaltung der Flexion. Die Version *Lemma+4POS* schließlich eliminiert durch die Kombination der Lemmata mit dem kleinen Tagset sowohl Flexion als auch die gegenüber Lemmata ohne solche Ergänzung verbleibenden Homographien.

Korpusversion	**Beispiel 1**	**wird zu**	**Beispiel 2**	**wird zu**
Lemmatisiert	Häusern	Haus	kauft ab	abkaufen
Wortform+STTS	Häusern	Häusern+NN	kauft ab	kauft+VVFIN
Lemma+STTS	Häusern	Haus+NN	kauft ab	abkaufen+VVFIN
Wortform+Lemma	Häusern	Häusern+Haus	kauft ab	kauft+abkaufen
Ohne Konjugation	Häusern	Häusern	kauft ab	abkaufen
Ohne Deklination	Häusern	Haus	kauft ab	kauft ab
Ohne PTKVZ	Häusern	Häusern	kauft ab	abkauft
Lemma+4POS	Häusern	Haus+S	kauft ab	abkaufen+S
Ohne Homographen	Häusern	Haus	Weine	weine+S/weine+V

Tabelle 5.4.: Modifizierte Korpusversionen mit Beispielen

Die beschriebenen Modifikationen vergrößern oder verkleinern die Menge der Types und damit die der potenziellen Klassifiktionsmerkmale gegenüber dem Originalkorpus durch

zwei verschiedene Mechanismen: Aufspaltung eines Terminalsymbols, also eines Types, in mehrere verschiedene durch das Hinzufügen von POS-Tags oder Lemmata, oder Fusionierung mehrerer Terminalsymbole in eines, etwa durch (Teil-)Lemmatisierung. Beide Mechanismen können alleinstehend, etwa in der Korpusversion *Lemmatisiert* (Fusionierung) oder *Ohne Homographen* (Aufspaltung), wirken oder konterkarierend, etwa in den Versionen *Lemma+STTS* und *Lemma+4POS*. Tabelle 5.5 zeigt die Anzahl der Types in den modifizierten Korpusversionen im Vergleich zur unlemmatisierten Originalversion.

Korpusversion	Kürzel	Anzahl Types
Wortform+STTS	WF+P	156.102
Wortform+Lemma	WF+L	155.021
Ohne PTKVZ	OP	147.100
Unlemmatisiert	**UL**	**146.423**
Ohne Homographen	OH	145.078
Ohne Konjugation	OK	139.268
Lemma+STTS	L+P	127.965
Ohne Deklination	OD	118.497
Lemma+4POS	L+4P	114.225
Lemmatisiert	L	113.615

Tabelle 5.5.: Anzahl der Types in den modifizierten Korpusversionen

Aus der Verteilung der Typesmengen wird das Interagieren der beiden widersprüchlichen Mechanismen deutlich: Zwei der drei größeren Typesmengen entstehen durch das Aufspalten der Types durch die POS-Tags oder verschiedene Lemmata. Als Beispiel für Ersteres gelte der Type *verrückt*, kontextabhängig ergänzt durch die STTS-Tags VVFIN, VVPP und ADJD, als Beispiel für Letzeres entsprechend die Ergänzung durch die Lemmata *verrücken* (zu VVFIN und VVPP) oder *verrückt* (ADJD). Der Umstand, dass es sich bei der dritten vergrößerten Typesmenge um die Korpusversion *Ohne PTKVZ* handelt, erklärt sich dadurch, dass aus einer Reihe von Verben wie *kauft ab* ein Type wie *abkauft* gebildet wird, während das Stammverb mitunter nichtsdestotrotz separat im Korpus auftritt und somit ein zusätzlicher Type ohne Verlust des Stammverb-Types erzeugt wird.

Die Gruppe der Korpusversionen mit reduzierter Typesmenge ist zu unterscheiden in die lemmatabasierten Korpusversionen sowie die drei phänomenreduzierenden Korpusversionen. Erstere reduzieren nachvollziehbarer Weise die Symbolmenge durch das Zusammenziehfen diverser Types in ihre Zitierform, am stärksten ohne Hinzufügung von POS-Tags in der ausschließlich lemmatisierten Version. Hingegen fällt neben den ebenfalls offensichtlich durch Fusionierung reduzierenden Korpusversionen *Ohne Deklination* und *Ohne Konjugation* die Version ohne Homographen ins Auge, da sie durch das Hinzufügen des kleinen POS-Tagsets eigentlich Terminalsymbole aufspaltet und die Typesmenge gegenüber der Originalversion vergrößert. Die laut dieser Aufstellung verringerte Typesmenge hat den technischen Grund, dass wie in Kapitel 2 erklärt die Groß- und Kleinschreibung als syntaktisch bestimmtes Wortmerkmal im Bag-of-Words-Modell entfällt und von den gängigen Implementierungen der Klassifikationsalgorithmen (inklusive dem verwendeten Scikit-learn) entfernt wird. Durch das Zusammenfallen von Terminalsymbolen wie *Wüste* (NN) und *wüste* (ADJD) oder *Weine* (VVIMP am Satzanfang), *weine* (VVFIN Satzmitte) und *Weine* (NN) wird in diese Korpusversion offensichtlich das Hinzufügen der Symbole des kleinen Tagsets überkompensiert.

Tabelle 5.6 summiert die Fähigkeiten der neun Korpusversionen, die vier Flexionsphänomene zu eliminieren, das heißt gänzlich aus dem Trainings- und Klassifikationsprozess zu entfernen, zu reduzieren oder zu erhöhen (ohne quantitative Aussage) oder unverändert zu lassen.

Klar ersichtlich korrespondieren hierbei die Korpusversionen *Wortform+STTS* und *Ohne Homographen* (äquivalent zu *Wortform+4POS*) einerseits sowie *Lemma+STTS* und *Lemma+4POS* andererseits hinsichtlich ihrer Reduktionsfähigkeiten. Bei formal gleichstarker Reduktionsfähigkeit, aber unterschiedlich großer Types-Menge erzielt die Korpusversion mit dem kleineren Tagset experimentell möglicherweise bessere Ergebnisse. Diese Vermutung ergibt sich aus der vorstehend getroffenen Annahme, dass stilistische Merkmale, wie sie durch das STTS zusätzlich abgebildet werden können, für eine The-

Korpusversion	Deklination	Konjugation	PTKVZ	Homographie
Lemmatisiert	**eliminiert**	**eliminiert**	**eliminiert**	reduziert
Wortform+STTS	unverändert	unverändert	unverändert	**eliminiert**
Lemma+STTS	**eliminiert**	**eliminiert**	**eliminiert**	**eliminiert**
Wortform+Lemma	unverändert	unverändert	**eliminiert**	**eliminiert**
Ohne Konjugation	unverändert	**eliminiert**	**eliminiert**	erhöht
Ohne Deklination	**eliminiert**	unverändert	unverändert	reduziert
Ohne PTKVZ	unverändert	unverändert	**eliminiert**	reduziert
Lemma+4POS	**eliminiert**	**eliminiert**	**eliminiert**	**eliminiert**
Ohne Homographen	unverändert	unverändert	unverändert	**eliminiert**

Tabelle 5.6.: Elimination Flexionsphänomene in verschiedenen Korpusversionen

menklassifikation irrelevant sind und das reduzierte Tagset statistisch aussagekräftigere Symbole mit deutlicherer Korrelation zur jeweiligen Klasse erzeugt.

Drei Korpusversionen beseitigen lediglich ein einziges Phänomen vollständig: *Ohne Deklination*, *Ohne PTKVZ* und *Ohne Homographen* eliminieren gerade das namensgebende Flexionsphänomen. Die Lemmatisierung der Verbformen im Korpus *Ohne Konjugation* löst zusätzlich redundant das Problem der abgetrennten Verbpartikeln (*kaufte ab – abkaufen*). Bei der Version *Ohne PTKVZ* handelt es sich somit im Grunde um eine Art Untermenge von *Ohne Konjugation* (*kaufte ab, kaufst ab – abkaufte, abkaufst*).

Sämtliche POS-Tags verwendenden Korpusversionen eliminieren das Phänomen der wortklassenübergreifende Homographie (Beispiel: *weine+VVFIN* vs. *weine+NN* oder *weine+V* vs. *weine+S*), während reine Lemmatisierung es etwa durch die Verringerung der Anzahl Types mit dem homographieanfälligen Suffix *-en* immerhin verringert. Hierbei wirken die Verbformen mit einer Zunahme der Types auf *-en* (*regten – regen*, *wagst – wagen*, *ruft - rufen*) und die quantitativ stärkere Verringerung der Types mit diesem Suffix bei den Substantiven (*weinen – wein*) gegeneinander. Neben der Endung *-en* sind wie in Unterkapitel 4.5 zur Homographie bereits dargestellt auch die Suffixe *-ern* (*modert – modern*, VVFIN vs. ADJD) und *-eln* (*deckelte – deckeln*, VVFIN vs. NN) in diesem

Zusammenhang für eine Zunahme der Homographen ausschließlich unter den Types der Korpusversion *Ohne Konjugation* verantwortlich.

Tabelle 5.7 konvertiert die Darstellung von Tabelle 5.6 in eine absteigend geordnete Übersicht zur Reduktionsfähigkeit der einzelnen Korpusversionen.

Korpusversion	**Basis**	**Eliminiert**	**Reduziert**	**Unverändert**	**E oder R***
Lemma+4POS	L*	4	0	0	4
Lemma+STTS	L	4	0	0	4
Lemmatisiert	L	3	1	0	4
Ohne Konjugation	WF*	2	1	1	3
Wortform+Lemma	WF	2	0	2	2
Ohne Deklination	WF	1	1	2	2
Ohne PTKVZ	WF	1	1	2	2
Ohne Homographen	WF	1	0	3	1
Wortform+STTS	WF	1	0	3	1

Tabelle 5.7.: Vergleich Reduktionsfähigkeiten der verschiedenen Korpusversionen, E oder R = Eliminiert oder reduziert, L = Lemmatabasiert, WF=Wortformenbasiert

Unmittelbar ersichtlich ist die Dominanz der lemmatabasierten Korpusversionen, unterbrochen nur von der vom vorstehend benannten Redundanzeffekt profitierenden konjugationsbefreiten Korpusversion. Die beiden Deklination und abtrennbare Verbpartikeln entfernenden Korpusversionen reduzieren neben der Beseitigung des namensgebenden Phänomens die Anzahl der Homographen: Das deklinationsfreie Korpus etwa über die Entfernung des homographieanfälligen Pluralsuffixes *-en* (Beispiele siehe oben) und das verbpartikelnfreie Korpus über die Anfügung der abgetrennten Partikeln an ihr Finitum (etwa *regte ab*, *regt ab* zu *abregen* (VVFIN) vs. *regen* (ambig: NN/VVINF/VVFIN)). Die beiden letztgenannten Korpusversionen, *Ohne Homographen* und *Wortform+STTS*, schneiden mit der ausschließlichen Beseitigung von Homographie am schwächsten ab. Hierbei steht zu erwarten, dass die Korpora, die STTS-Tags an die unlemmatisierten Wortformen anfügen, aus oben genannten Gründen experimentell schwächer abschneiden werden als *Ohne Homographen* bei Verwendung des kleinen POS-Tagsets.

Tabelle 5.8 wechselt die Perspektive und zeigt auf, in wievielen der neun modifizierten Korpusvarianten die drei ursächlichen und das symptomatische Flexionsphänomen eliminiert oder reduziert werden können.

Abbau	**Deklination**	**Konjugation**	**PTKVZ**	**Homographie**
Eliminiert	4	4	6	5
Reduziert	0	0	0	4
Unverändert	5	5	3	0
E oder R	4	4	6	9

Tabelle 5.8.: Eliminierbarkeit der einzelnen Phänomene in verschiedenen Korpusversionen

Unmittelbar ersichtlich werden Deklination und Konjugation neben der jeweiligen spezialisierten Korpusversion in den drei ausschließlich lemmatabasierten Korpusversionen eliminiert. Die Korpusversion *Wortform+Lemma* leistet hier keine guten Dienste, da sie in diesem Zusammenhang Types ausdifferenziert, aber nicht zusammenzieht (etwa *Häusern+Haus* oder *tranken+trinken*. Das Phänomen der abtrennbaren Verbpartikeln wird zusätzlich wie vorstehend beschrieben durch das konjugationsbefreite Korpus aufgelöst und in der Version *Wortform+Lemma* um den Preis zusätzlich eingeführter Symbole disambiguiert, etwa *kaufte+abkaufen*. Das Phänomen der Homographie wird in der Mehrheit der Korpusversionen vollständig und in den verbleibenden Versionen teilweise beseitigt, da, wie in Unterkapitel 4.5 dargestellt, die weit überwiegende Mehrheit der Homographien auf gängigen Flexionsmorphemen beruht, die mit anteiliger oder vollständiger Lemmatisierung reduziert werden.

5.2.1.2. Zusammenfassung und Diskussion

Die vorstehenden Betrachtungen zeigen, dass sowohl die zu untersuchenden Flexionsphänomene als auch die zur Untersuchung herangezogenen Korpusversionen potenziell unterschiedlich effektiv sein können: Der Anteil der neun Korpora, die ein spezifisches Phä-

nomen der Flexionsmorphologie eliminieren oder reduzieren können, reicht von vier bei Deklination und Konjugation bis zu sämtlichen neun bei Homographie. Aus umgekehrter Perspektive reicht die Eliminationsfähigkeit der Korpora auf die Flexionsphänomene bezogen von vollständig bis ausschließlich homogaphiebezogen (lemmatabasiert vs. wortformenbasiert, s. Tabelle 5.7).

Eine Festlegung, welche Flexionsphänomene zum einen im vorliegenden Klassifikationsszenario den stärksten Einfluss auf den Klassifikationserfolg ausüben, und mit welchen korpusmodifizierenden und somit merkmalsraumgestaltenden Maßnahmen ihnen zum anderen beizukommen ist, kann letztlich nur empirisch getroffen werden (s. Abschnitt 5.2.2).

Aus den Überlegungen zu den Einflüssen von Deklination, Konjugation, abtrennbaren Verbpartikeln und Homographie in Unterkapitel 4.6 und der empirischen Analyse der Entwicklung der Merkmalsräume im vorhergehenden Unterkapitel wurde die Vermutung begründet, dass die Deklination durch die dominante Stellung der Substantive im Merkmalsraum mindestens anfänglich, das heißt in kleineren Merkmalsräumen, den größten Einfluss auf den Klassifikationserfolg ausübt. Mit dem leicht zunehmenden Anteil der Verben mit Vergrößerung des Merkmalsraums stünde ein größerer Einfluss der Konjugation im Merkmalsraum zu erwarten. Dessen Auswirkungen auf den Klassifikationsprozess ingesamt wiederum hängen an der Korrelation zwischen Merkmalsraumgröße und Klassifikationserfolg insgesamt: Bietet ein erweiterter Merkmalsraum Erfolgspotenzial für das Klassifikationsergebnis, sind Gewinne oder Verluste hierbei von der Einwirkung und somit Kompensation der wortklassenbasierten Flexionsphänomene entscheidend für die insgesamt erreichbare Klassifikationsleistung.

Die Entwicklungen der Merkmalsräume, wie in Unterkapitel 5.1 dokumentiert, zugrundelegend, kann aus den vorstehenden Korpusbetrachtungen die Überlegung abgeleitet werden, dass deklinationsfreie Korpusversionen anfänglich den größeren Klassifikations-

erfolg bieten sollten, im Laufe der Merkmalsraumvergrößerung jedoch von Korpusversionen abgelöst werden, die zusätzliche Phänomene ergänzend abdecken, wobei wiederum Versionen mit geringerer Typesmenge formal gleich starken mit größerer Typesmenge vorzuziehen sind. Hierbei handelt es sich aller Voraussicht nach somit um die lemmatabasierte, das reduzierte POS-Tag-Set verwendende Korpusversion.

5.2.2. Ergebnisse und Analyse der konventionellen Klassifikationsexperimente

Dieser Abschnitt dokumentiert die Ergebnisse von sechs konventionellen Klassifikatoren auf dem Originalkorpus sowie auf den neun in Abschnitt 5.2.1 beschriebenen modifizierten Korpusversionen. Die Scikit-learn-Implementationen der verwendeten Klassifikationsverfahren K-nächste-Nachbarn, Logistische Regression, Naive Bayes (Binomial und multinomial), Rocchio und Support Vector Machine entsprechen den Beschreibungen in Unterkapitel 4.2 und wurden ohne abweichende Parametrisierungen der Standardimplementation entnommen. Jeder Klassifikator wurde in jeder der Testreihen auf einer in Schritten von fünf Texten anwachsenden Trainingsmenge, beginnend bei fünf und endend bei 3.600 Texten (720 Schritte), trainiert. Dem Klassifikator wurden in jeder Testreihe die bereits beschriebenen 13 verschiedenen Merkmalsraumgrößen, von zehn Merkmalen bis zu sämtlichen Merkmalen, zur Verfügung gestellt. Somit ergaben sich pro Korpusversion und Klassifikator 93.600 zu trainierende Modelle, die zur Kreuzvalidierung je acht Mal erzeugt und getestet wurden. Sämtliche Kombinationen aus Klassifikator, Trainingsmenge, Merkmalsraumgröße und Korpusversion ergaben somit eine Menge von 748.800 einzelnen Experimenten. Der Test jedes einzelnen erzeugten Modells erfolgte auf stets 100 aus dem verbleibenden verfügbaren Korpus zufallsgesampelten unbekannten Texten. Zur weiteren Erhöhung der Varianz der Trainingsmenge wurde diese für eine neue Trainingsmenge nicht jeweils durch Hinzufügen neuer Zufallstexte zur letz-

ten Trainingsmenge erzeugt, sondern durch vollständig zufälliges Neuzusammenstellen in jedem Vergrößerungsschritt.

In den folgenden Unterabschnitten 5.2.2.1 bis 5.2.2.6 werden die Ergebnisse der Klassifikationsverfahren im beschriebenen Experimentalraum kompakt in Form von graphischen wie tabellarischen Übersichten präsentiert. Eine Kommentierung in Textform erfolgt nur, soweit sie der Vorbereitung auf die ausführlichere Gesamtanalyse in Unterabschnit 5.2.2.7 dient.

Für die visuelle Darstellung der Ergebnisse wurde die Skalierung der Y-Achse für den relevanten Ausschnitt der jeweiligen F-Scores angepasst. Zusätzlich zu beachten ist zudem die aus Übersichtlichkeitsgründen nicht skalierte Schrittweite der Merkmalsraumvergrößerungen, insbesondere im Hinblick auf sichtbare Konvergenzentwicklungen.

Die folgende kompakte Übersicht deckt sich mit der in der Literatur etablierten Feststellung, dass die Leistung etablierter Klassifikationsalgorithmen hochgradig aufgabenspezifisch ausfällt und sich innerhalb desselben Szenarios erheblich unterscheiden kann. Diese Regel trifft erkennbar auch im vorliegenden Szenario zu: Bei vollständig gleichem Experimentalszenario verhalten sich die Klassifikatoren auf demselben Korpus in Hinsicht auf absolute Leistungsfähigkeit nach F-Score, Lerngeschwindigkeit, Konvergenzverhalten, Fluktuation und Volatilität gegenüber den Korpusmodifikationen stark unterschiedlich.

Während die Klassifkatoren *Binomial Naive Bayes*, *Logistische Regression* und *Support Vector Machine* in den jeweils informationsoptimierten Korpusversionen in großen Merkmalsräumen F-Scores stabil oberhalb von 0,6 erreichen, zeigen sich der *K-nächste-Nachbarn-Klassifikator* und der konzeptionell ähnliche *Rocchio-Klassifikator* mit absolut schwächeren F-Scores sowie starker Volatilität gegenüber größeren Merkmalsräumen. Während der *Rocchio-Klassifikator* bauartbedingt möglicherweise langfristig von sich konsolidierenden Merkmalswerten durch deren sinkende Varianz stabilisiert wird, ist im

Experiment eine solche Tendenz beim *K-nächste-Nachbarn-Klassifikator* nicht abzusehen. Das multinomiale Modell des *Naive Bayes-Klassifikators* hingegen erreicht der ersten Gruppe zeitweise ebenbürtige F-Score-Größenordnungen, vollzieht aber eine spätere und höher angesiedelte analoge Entwicklung zum *K-nächste-Nachbarn-Klassifikator* in Form von einsetzendem Leistungsverlust.

Die vorstehenden Anmerkungen beschränken sich bewusst weitgehend auf Beobachtungen, die sich auf die absolute Leistungsfähigkeit der Klassifikatoren und die Entwicklung der Merkmalsraumgrößungen beziehen. Das im Zentrum dieser Untersuchung stehende Verhalten der Klassifikatoren gegenüber den flexionsreduzierenden Modifikationen des Korpus ist Gegenstand der Gesamtanalyse in Unterabschnitt 5.2.2.7.

Die Präsentation der vier nach Leistungsfähigkeit und Konvergenzverhalten ähnlichen Klassifikatoren erfolgt im folgenden Format: Einer Tabelle zur übersichtlichen Darstellung der Erstplatzierungen nach Korpusversion pro Merkmalsraum und zur Herausstellung der besten erreichten F-Scores folgt eine Visualisierung der Entwicklung des F-Scores bei wachsenden Merkmalsräumen. Die jeweils besten Ergebnisse in einer Merkmalsraumgröße wurden einem Zwei-Stichproben-T-Test im Vergleich mit der unlemmatisierten Korpusversion unterzogen, dessen Ergebnis unterhalb der Ergebnisse in Blautönen zu $p < 0,05$ und $p < 0,10$ angezeigt wird. Die Erstplatzierungen in der erstgenannten Tabelle sind unabhängig von der absoluten Ergebnishöhe grün markiert, die zweite Tabelle zeigt F-Scores von 0,60 oder besser in hellgrün, F-Scores von 0,65 oder besser dunkelgrün unterlegt. Die Tabelle mit der Markierung der Erstplatzierungen ist um zusammenfassende Zeilen zu merkmalsraumbezogenen Minima, Maxima und deren Differenz ergänzt. Für die Logistische Regression stellvertretend für die oben benannte Vierergruppe wird zudem eine Übersicht über Lernverhalten und Konvergenztendenzen im Hinblick auf die Größe der Trainingsmenge und Merkmalsräume präsentiert.

Das Darstellungsformat entspricht den in Unterkapitel 2.3 formulierten verschiedenen Perspektiven in der Evaluation eines Klassifikators. Neben dem Ziel, das beste erreichbare Ergebnis in Form eines (gewichteten) F-Scores zu generieren, können weitere Interpretationen als Nebenziele oder Alternativen zur Beurteilung eines Klassifikationsmodells gewählt werden. Unter diesen sind der Wunsch nach zügiger Konvergenz, interpretierbar als effektiver Umgang mit restringierten Trainingsmengen, und die Garantie eines Mindestscores zugunsten vereinzelt auftretender Leistungsspitzen in den folgenden Darstellungen wiederzufinden. Die ebenso denkbare Zielsetzung kohärenter, klassenübergreifend gleichmäßiger Leistungscharakteristika, steuerbar durch Mikro- oder Makro-Gewichtung, wird in dieser Untersuchung hingegen als primär semantisch aufgefasst nicht weiter behandelt.

5.2.2.1. K-nächste-Nachbarn und Rocchio-Nearest Centroid

Abbildung 5.10 zeigt die Charakteristika des in diesem Klassifikationsszenario nicht konkurrenzfähigen und daher nicht weiter diskutierten K-nächste-Nachbarn-Klassifikators. Der Rocchio-Klassifikator zeigt ausweislich Abbildung 5.11 eine ähnlich niedrige Klassifikationsleistung wie der K-nächste-Nachbarn-Klassifikator und wird daher ebenfalls von der weiteren Diskussion ausgenommen.

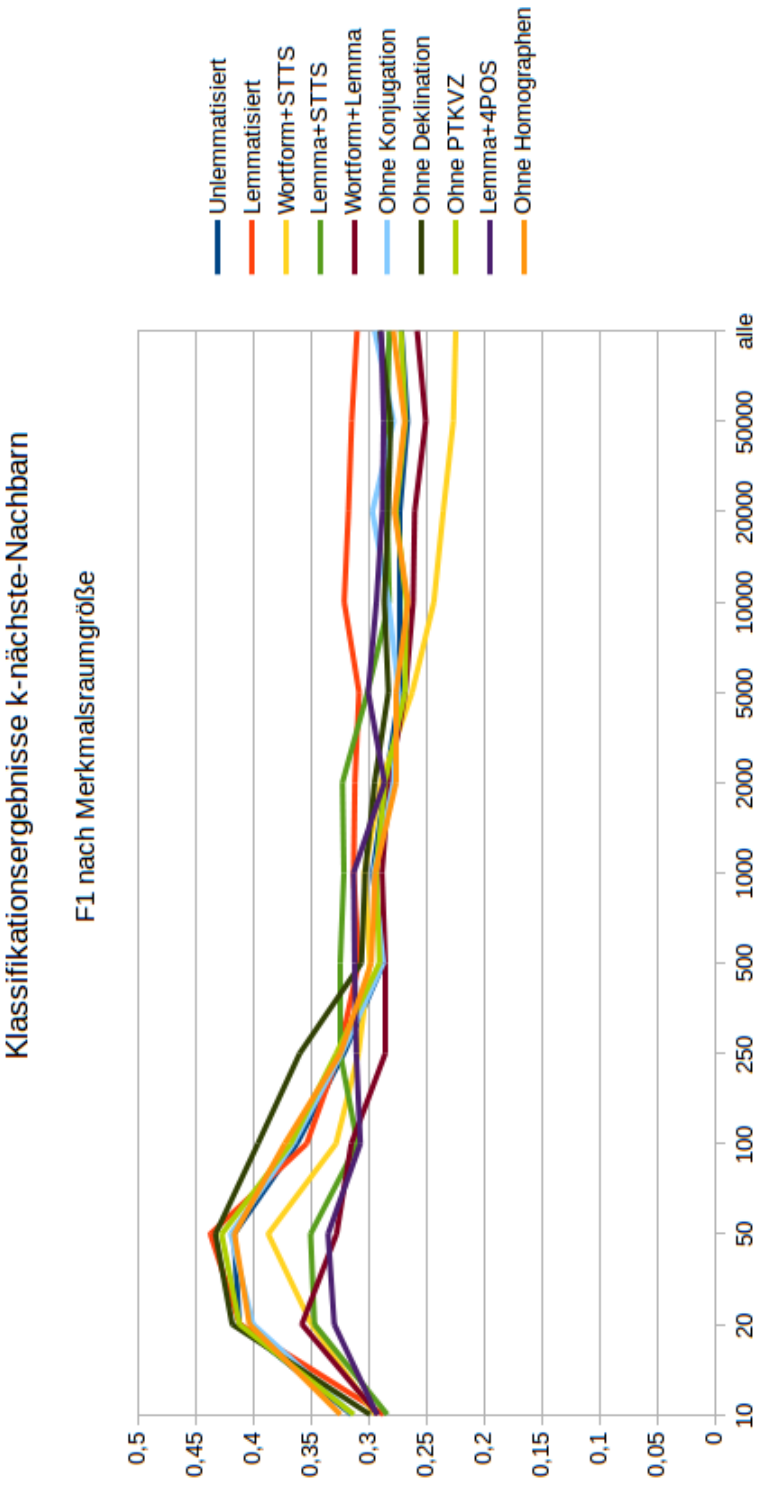


Abbildung 5.10.: Klassifikationsergebnisse K-nächste-Nachbarn

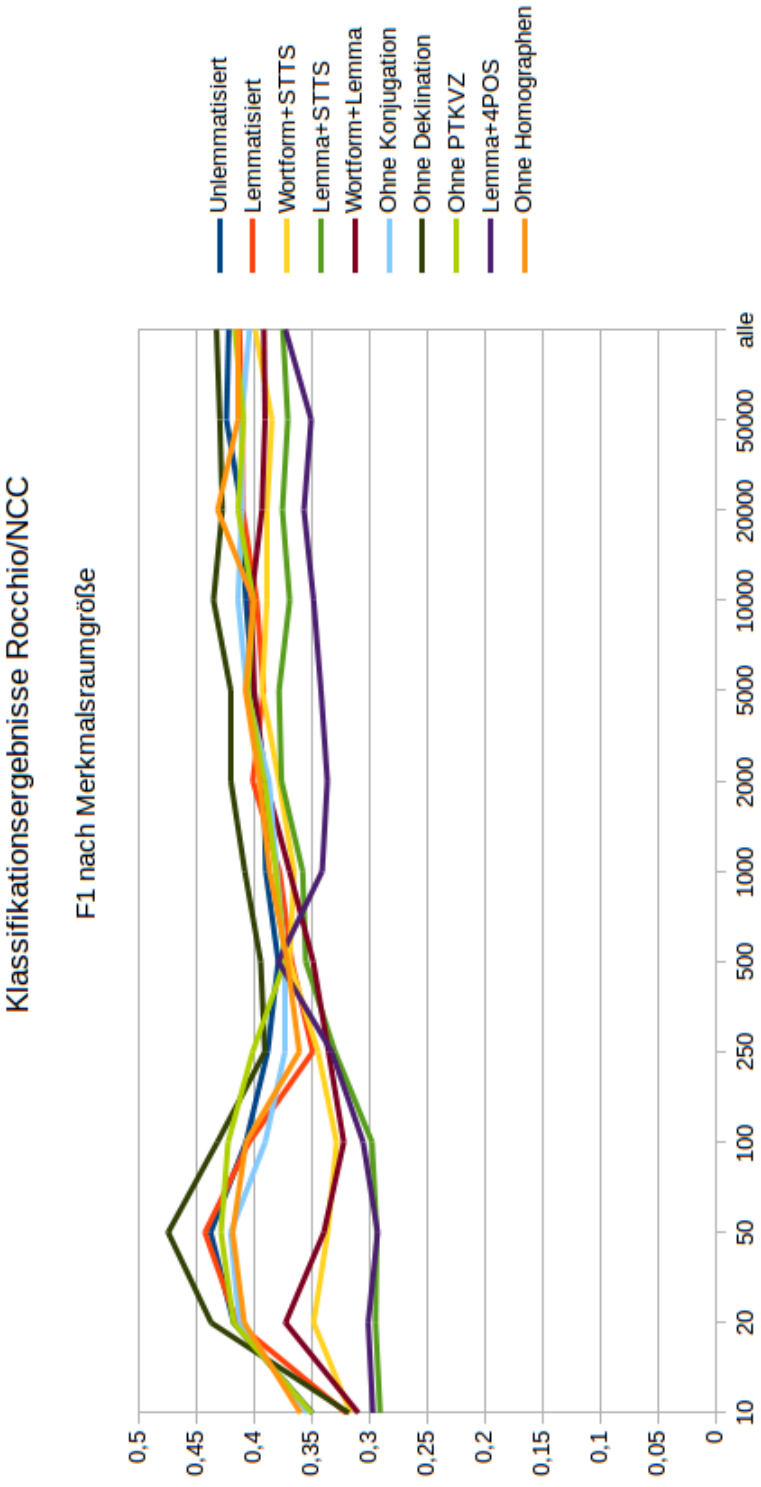
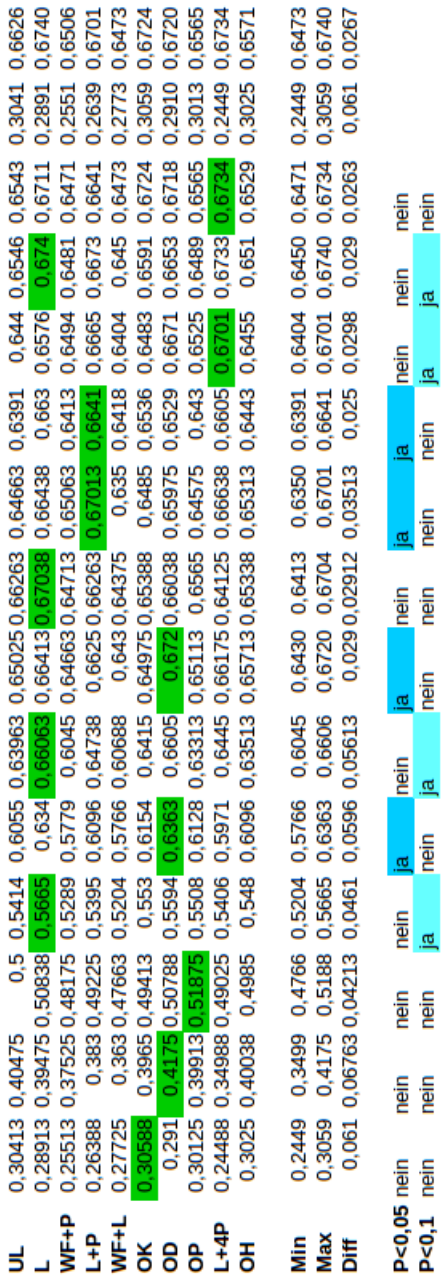


Abbildung 5.11.: Klassifikationsergebnisse Rocchio

5.2.2.2. Logistische Regression

Die Abbildungen 5.12 bis 5.14 zeigen die experimentellen Ergebnisse der Klassifikation mittels logistischer Regression. Dieser Klassifikator erzielt mit einem F-Score von 0,674 bei 50.000 Merkmalen auf dem lemmatisierten Korpus das beste Ergebnis aller konventionellen Experimente. Zum Zeitpunkt der Heranziehung sämtlicher Merkmale kann keine sichere Aussage über einsetzende Konvergenz getroffen werden: Vier der Korpusversionen erzeugen weiterhin (leicht) steigende F-Scores, fünf, darunter das erstplatzierte Szenario *Lemmatisiert* bereits einen leichten Abfall. Die beiden schlechtesten Maximalergebnisse werden von den wortformbasierten Korpusversionen *Wortform+STTS* und *Wortform+Lemma* erzeugt. Die auf die vorliegende Gesamthöhe insgesamt erreichbare Verbesserung von 1,14 F-Score-Punkten durch die lemmatisierte Korpusversion erscheint gering; nur in je vier Merkmalsräumen kann eine stark beziehungsweise moderat signifikante Verbesserung durch Korpusmodifikationen erzielt werden. Der Verlauf der Differenz zwischen Minimal- und Maximalwert pro Merkmalsraumgröße zeigt ein auch bei den weiteren besprochenen Klassifikatoren zu beobachtendes Muster: Die erreichbare Verbesserung durch Korpusmodifikationen sinkt von eingangs mehr als 6 Punkten langfristig auf 2,63 Punkte bei Verwendung aller Merkmale. Die Tendenz, dass die Größe des Merkmalsraums die Entwicklung der F-Scores ungleich stärker beeinflusst als die Modifikationen am Korpus, ist daher zusätzlicher Gegenstand der anschließenden Gesamtschau.



	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,30413	0,40475	0,5	0,5414	0,6055	0,63963	0,65025	0,66263	0,64663	0,6391	0,644	0,6546	0,6543	0,3041	0,6626
L	0,28913	0,39475	0,50838	0,5665	0,634	0,66063	0,66413	0,67038	0,66438	0,663	0,6576	0,674	0,6711	0,2891	0,6740
WF+P	0,25513	0,37525	0,48175	0,5289	0,5779	0,6045	0,64663	0,64713	0,65063	0,6413	0,6494	0,6481	0,6471	0,2551	0,6506
L+P	0,26388	0,383	0,49225	0,5395	0,6096	0,64738	0,6625	0,66263	0,67013	0,6641	0,6665	0,6673	0,6641	0,2639	0,6701
WF+L	0,27725	0,363	0,47663	0,5204	0,5766	0,60688	0,643	0,64375	0,635	0,6418	0,6404	0,645	0,6473	0,2773	0,6473
OK	0,30588	0,3965	0,49413	0,553	0,6154	0,6415	0,64975	0,65388	0,6485	0,6536	0,6483	0,6591	0,6724	0,3059	0,6724
OD	0,291	0,4175	0,50788	0,5594	0,6363	0,6605	0,672	0,66038	0,65975	0,6529	0,6671	0,6653	0,6718	0,2910	0,6720
OP	0,30125	0,39913	0,51875	0,5508	0,6128	0,63313	0,65113	0,6565	0,64575	0,643	0,6525	0,6489	0,6565	0,3013	0,6565
L+4P	0,24488	0,34988	0,49025	0,5406	0,5971	0,6445	0,66175	0,64125	0,66638	0,6605	0,6701	0,6733	0,6734	0,2449	0,6734
OH	0,3025	0,40038	0,4985	0,548	0,6096	0,63513	0,65713	0,65338	0,65313	0,6443	0,6455	0,651	0,6529	0,3025	0,6571
Min	0,2449	0,3499	0,4766	0,5204	0,5766	0,6045	0,6430	0,6413	0,6350	0,6391	0,6404	0,6450	0,6471	0,2449	0,6473
Max	0,3059	0,4175	0,5188	0,5665	0,6363	0,6606	0,6720	0,6704	0,6701	0,6641	0,6701	0,6740	0,6734	0,3059	0,6740
Diff	0,061	0,06763	0,04213	0,0461	0,0596	0,05613	0,029	0,02912	0,03513	0,025	0,0298	0,029	0,0263	0,061	0,0267
P<0,05	nein	nein	nein	nein	ja	nein	ja	nein	ja	ja	nein	nein	nein		
P<0,1	nein	nein	nein	ja	nein	ja	nein	nein	nein	nein	ja	ja	nein		

Abbildung 5.12.: Klassifikationsergebnisse Logistische Regression, tabellarisch (Erstplatzierungen und P-Werte markiert)

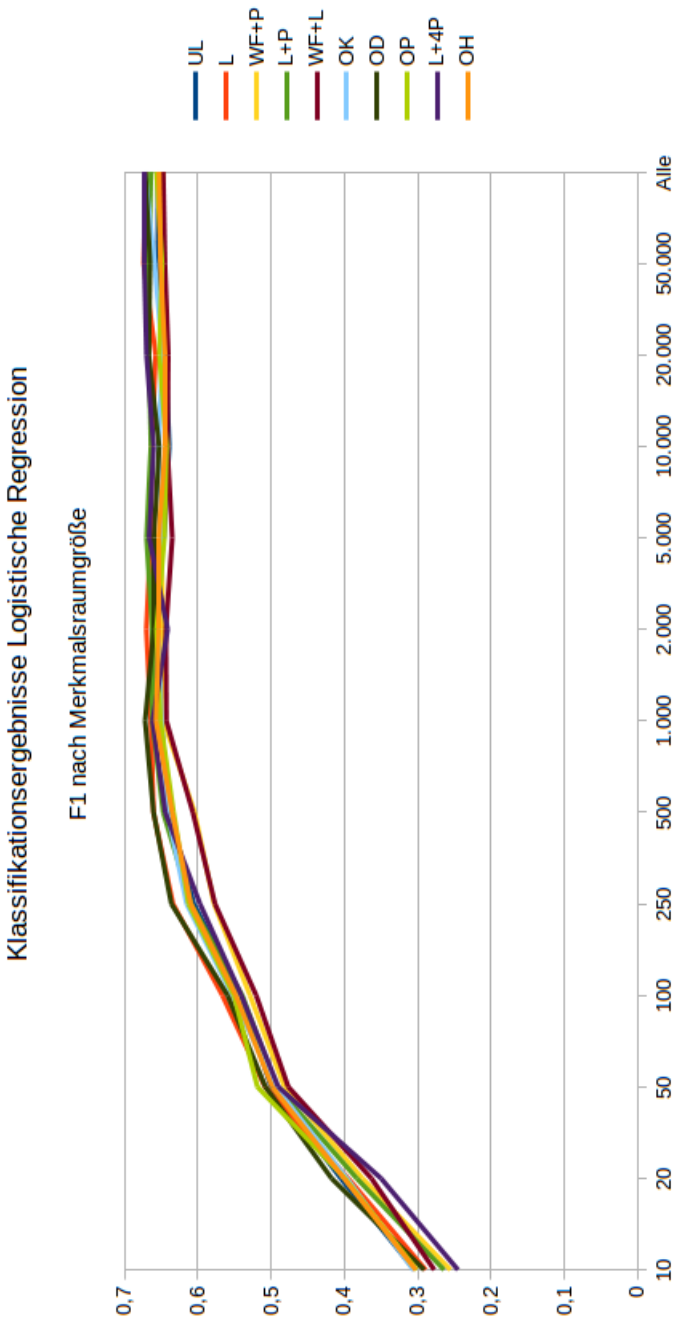


Abbildung 5.13.: Klassifikationsergebnisse Logistische Regression - F1 nach Merkmalsraumgröße

Abbildung 5.12 zeigt ein uneinheitliches Bild der Verteilung der Erstplatzierungen nach Korpusmodifikationen. Lemmatabasierte Korpusversionen mit und ohne POS-Tags erreichen mit 8 Erstplatzierungen eine absolute Mehrheit. *Ohne Deklination* stellt als darauf folgende stärkste Einzelmodifikation zwei der vier statistisch hoch signifikanten Erstplatzierungen gleichauf mit *Lemma+POS-Tag* mit ebenfalls zwei hoch signifikanten Verbesserungen. Tendenziell dominiert die Modifikation *Ohne Deklination* Merkmalsräume bis 1.000 Merkmale, bevor sie von ausschließlich lemmatabasierten Korpusversionen abgelöst wird. Diese Entwicklung ist aufgrund ihres Auftretens auch bei den verbleibenden Klassifikatoren ebenfalls Gegenstand der Gesamtanalyse in Unterabschnitt 5.2.2.7.

	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,30413	0,40475	0,5	0,5414	0,6055	0,63963	0,65025	0,66263	0,64663	0,6391	0,644	0,6546	0,6543	0,3041	0,6626
L	0,28913	0,39475	0,50838	0,5665	0,634	0,66063	0,66413	0,67038	0,66438	0,663	0,6576	0,674	0,6711	0,2891	0,6740
WF+P	0,25513	0,37525	0,48175	0,5289	0,5779	0,6045	0,64663	0,64713	0,65063	0,6413	0,6494	0,6481	0,6471	0,2551	0,6506
L+P	0,26388	0,383	0,49225	0,5395	0,6096	0,64738	0,6625	0,66263	0,67013	0,6641	0,6665	0,6673	0,6641	0,2639	0,6701
WF+L	0,27725	0,363	0,47663	0,5204	0,5766	0,60688	0,643	0,64375	0,635	0,6418	0,6404	0,645	0,6473	0,2773	0,6473
OK	0,30588	0,3965	0,49413	0,553	0,6154	0,6415	0,64975	0,65388	0,6485	0,6536	0,6483	0,6591	0,6724	0,3059	0,6724
OD	0,291	0,4175	0,50788	0,5594	0,6363	0,6605	0,672	0,66038	0,65975	0,6529	0,6671	0,6653	0,6718	0,2910	0,6720
OP	0,30125	0,39913	0,51875	0,5508	0,6128	0,63313	0,65113	0,6565	0,64575	0,643	0,6525	0,6489	0,6565	0,3013	0,6565
L+4P	0,24488	0,34988	0,49025	0,5406	0,5971	0,6445	0,66175	0,64125	0,66638	0,6605	0,6701	0,6733	0,6734	0,2449	0,6734
OH	0,3025	0,40038	0,4985	0,548	0,6096	0,63513	0,65713	0,65338	0,65313	0,6443	0,6455	0,651	0,6529	0,3025	0,6571
Min	0,2449	0,3499	0,4766	0,5204	0,5766	0,6045	0,6430	0,6413	0,6350	0,6391	0,6404	0,6450	0,6471	0,2449	0,6473
Max	0,3059	0,4175	0,5188	0,5665	0,6363	0,6606	0,6720	0,6704	0,6701	0,6641	0,6701	0,6740	0,6734	0,3059	0,6740
Diff	0,061	0,06763	0,04213	0,0461	0,0596	0,05613	0,029	0,02912	0,03513	0,025	0,0298	0,029	0,0263	0,061	0,0267

Abbildung 5.14.: Klassifikationsergebnisse Logistische Regression, tabellarisch (F-Scores >0,60 markiert)

Abbildung 5.14 zeigt eine klar erkennbare Tendenz sowohl in Hinblick auf die Merkmalsraumgröße als auch die Zuordnung zu den Korpusversionen: F-Scores über 0,6 treten erstmals, und sogleich mit deutlicher Mehrheit, in den Merkmalsräumen ab 250 auf und sind ab 500 Merkmalen garantiert. Ab 500 Merkmalen treten erstmalig F-Scores von über 0,65 auf, woraufhin sie ab 1.000 Merkmalen umgehend die Hälfte bis zur überwiegenden Mehrheit (70% bei 50.000 Merkmalen und 80% bei sämtlichen Merkmalen) der Ergebnisse ausmachen.

Die Visualisierungen 5.15 bis 5.17 zeigen den Lernverlauf dreier ausgewählter Korpusversionen: Der unlemmatisierten Originalversion sowie der beiden erstplatzierten Versionen *Lemmatisiert* und *Ohne Deklination*. Die jeweils am rechten Ende der X-Achse und somit bei 100% verwendetem Trainingskorpus befindlichen Werte sind in den entsprechenden Zellen der Abbildungen 5.12 und 5.14 zu finden.

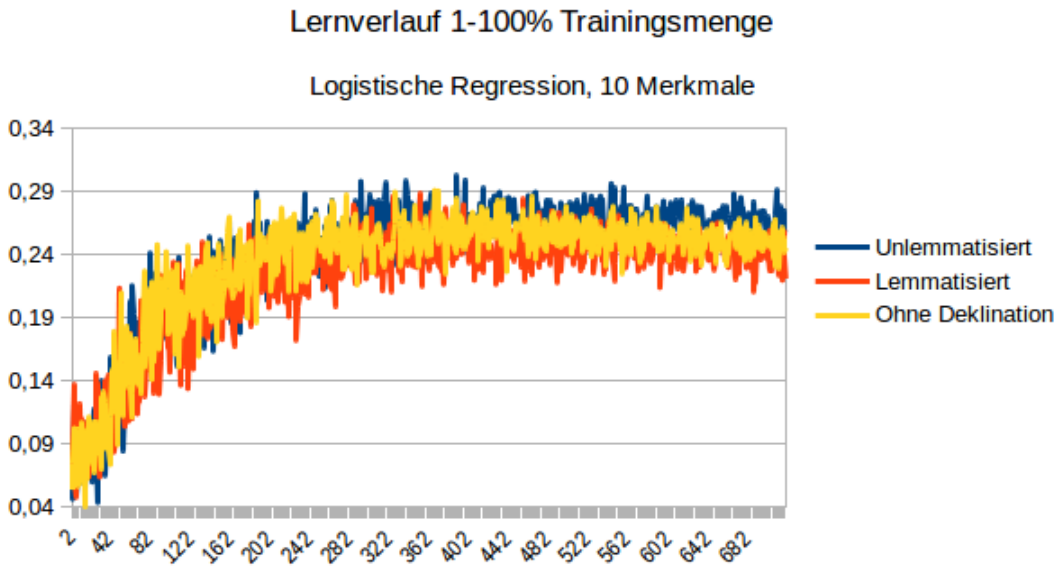


Abbildung 5.15.: Lernverlauf Logistische Regression, drei Korpusversionen, 10 Merkmale, 1-100% Trainingsmenge

Abbildung 5.15 zeigt typisches Verhalten eines im Vergleich zur potenziell als Merkmale zur Verfügung stehenden Typesmenge sehr kleinen Merkmalsraums in mehrfacher Hinsicht: Starke Fluktuation auch im anhaltenden Lernverlauf, eine untergeordnete Rolle der später dominierenden und leistungsfähigen Korpusversionen, sowie bereits visuell ersichtlich ausbleibende Konvergenz. Offensichtlich überschreitet die mit sämtlichen Trainingstexten unterbreitete Varianz im Korpus die Formalisierungsmöglichkeiten mit lediglich zehn Merkmalen sowohl der originalen als auch der modifizierten Korpusversionen für diesen Klassifikator unabhängig vom absolut ungünstigen F-Score. Der Bedarf an mehr als zehn oder wenigen dutzend Merkmalen zur Erstellung eines tragfähigen Sprachmodells in einem Klassifikationsszenario mit acht Kategorien erscheint nachvollziehbar und ist auch in den sonstigen Klassifikatoren der Vierergruppe zu beobachten. Die Abstände der abweichend platzierten Korpusversionen untereinander bleiben aufgrund der hohen Fluktuation unterhalb der Varianz innerhalb der Modelle.

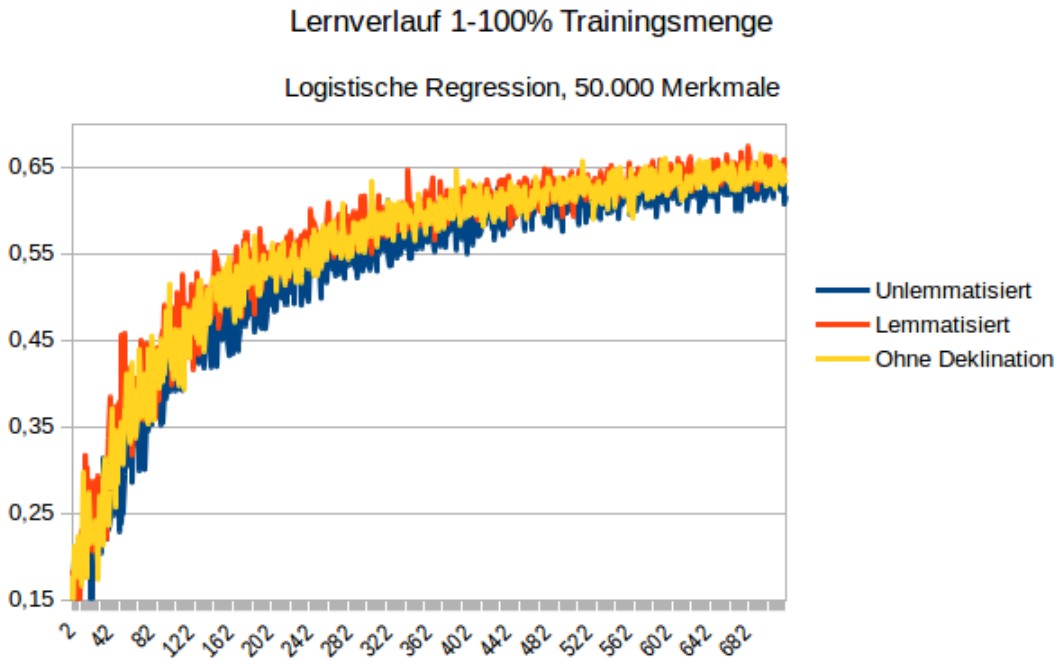


Abbildung 5.16.: Lernverlauf Logistische Regression, drei Korpusversionen, 50.000 Merkmale, 1-100% Trainingsmenge

Abbildung 5.16 zeigt dem Gesamtergebnis des Klassifikators und den Abbildungen 5.12 und 5.13 entnehmbares Verhalten bei 50.000 Merkmalen: Sichtbare Konvergenztendenz über den Trainingsverlauf, zügig einsetzend dem Endergebnis folgende Rangfolge der Korpusversionen sowie deutlich verringerte Fluktuation, erklärbar durch geringere Mobilität innerhalb des großen Merkmalsraums zwischen den Datenpunkten.

Die trainingsmengenbasierte Beobachtung des Lernverlaufs motiviert eine Ergänzung um eine Visualisierung der maximal erreichbaren Leistungsfähigkeit der Logistischen Regression im Hinblick auf die Merkmalsraumgröße.

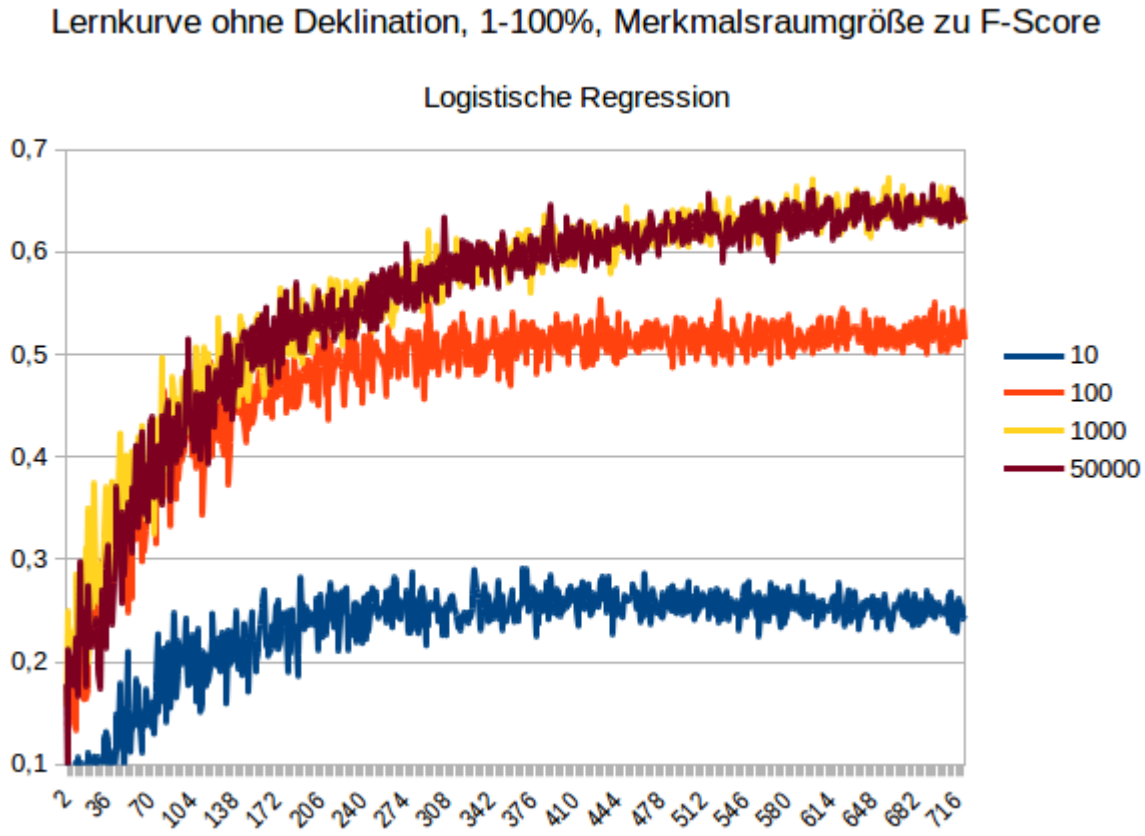
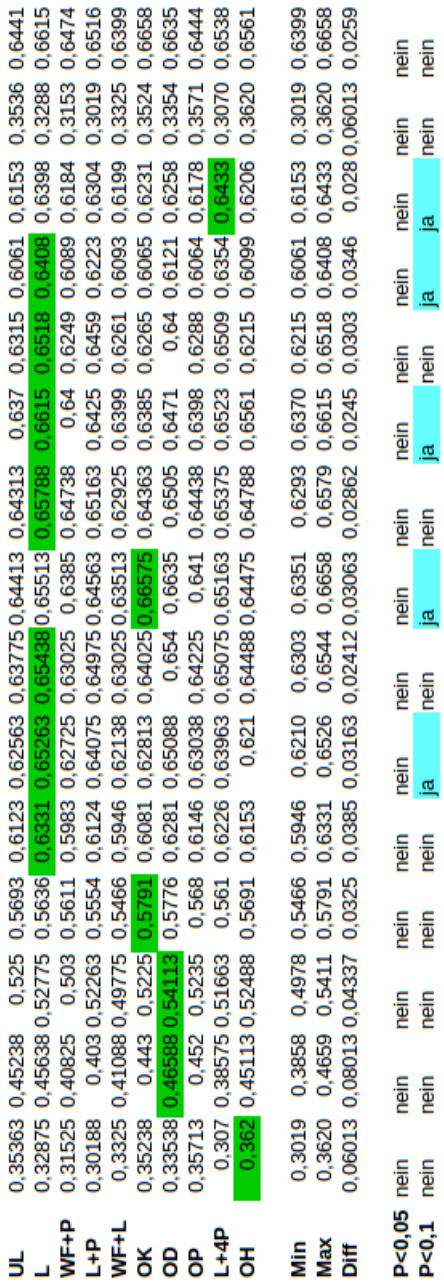


Abbildung 5.17.: Lernverlauf Logistische Regression, drei Korpusversionen, 50.000 Merkmale, 1-100% Trainingsmenge

Abbildung 5.17 zeigt das Konvergenzverhalten des Klassifikators exemplarisch für die leistungssteigernde Korpusmodifikation *Ohne Deklination*. Die Steigerung des F-Scores verlangsamt sich ebenso mit zunehmendem Trainingsmengenaufwand wie mit der Vergrößerung des Merkmalsraums: Folgt aus einer initialen Verzehnfachung der Merkmale im ersten Schritt noch eine Verdopplung des F-Scores, verringert sich dieser Gewinn bei der folgenden Verzehnfachung auf gute 20 Punkte, der keine weitere Steigerung bei 50.000 Merkmalen mehr folgt. Die auf 50.000 Merkmalen basierenden Modelle weisen eine flachere Lernkurve und somit langsamere Konvergenzentwicklung auf, bevor sie sich in der Größenordnung des sehr viel kleineren Merkmalsraumes von 1.000 Merkmalen einfinden. Ebenfalls erneut augenfällig ist eine deutlich abnehmende Fluktuation mit zunehmendem Trainingsfortschritt. Alle genannten Beobachtungen zeigen eine Konsolidierung des Modells, wobei Modelle mit größeren Merkmalsräumen langsamer konvergieren und des Weiteren ihre Ausdrucksstärke offensichtlich erst ab einem gewissen Mindesttrainingsstand ausspielen können; so erfolgt eine deutliche Abkopplung des Modells mit 100 Merkmalen zu dauerhaft niedriger Leistungsfähigkeit erst in der Umgebung von etwa 1.000 Trainingstexten.

5.2.2.3. Binomial Naive Bayes

Die Abbildungen 5.18 bis 5.20 präsentieren die Ergebnisse des mit einem F1 von 0,6558 zweitplatzierten konventionellen Klassifikators *Binomial Naive Bayes* mit einem in den kleineren Merkmalsräumen der Logistischen Regression ähnlichen Fluktuationsmuster in Bezug auf Erstplatzierungen und einem sich ebenfalls mutmaßlich stabilisierenden Konvergenzverhalten: Die Erstplatzierungen ab 250 Merkmalen, unterbrochen ein letztes Mal bei 2.000 Merkmalen, werden durchgehend von den lemmatabasierten Korpusversionen erreicht.



	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,35363	0,45238	0,525	0,5693	0,6123	0,62563	0,63775	0,64413	0,64313	0,637	0,6315	0,6061	0,6153	0,3536	0,6441
L	0,32875	0,45638	0,52775	0,5636	0,6331	0,65263	0,65438	0,65513	0,65788	0,6615	0,6518	0,6408	0,6398	0,3288	0,6615
WF+P	0,31525	0,40825	0,503	0,5611	0,5983	0,62725	0,63025	0,6385	0,64738	0,64	0,6249	0,6089	0,6184	0,3153	0,6474
L+P	0,30188	0,403	0,52263	0,5554	0,6124	0,64075	0,64975	0,64563	0,65163	0,6425	0,6459	0,6223	0,6304	0,3019	0,6516
WF+L	0,3325	0,41088	0,49775	0,5466	0,5946	0,62138	0,63025	0,63513	0,62925	0,6399	0,6261	0,6093	0,6199	0,3325	0,6399
OK	0,35238	0,443	0,5225	0,5791	0,6081	0,62813	0,64025	0,66575	0,64363	0,6385	0,6265	0,6065	0,6231	0,3524	0,6658
OD	0,33538	0,46588	0,54113	0,5776	0,6281	0,65088	0,654	0,6635	0,6505	0,6471	0,64	0,6121	0,6258	0,3354	0,6635
OP	0,35713	0,452	0,5235	0,568	0,6146	0,63038	0,64225	0,641	0,64438	0,6398	0,6288	0,6064	0,6178	0,3571	0,6444
L+4P	0,307	0,38575	0,51663	0,561	0,6226	0,63963	0,65075	0,65163	0,65375	0,6523	0,6509	0,6354	0,6433	0,3070	0,6538
OH	0,362	0,45113	0,52488	0,5691	0,6153	0,621	0,64488	0,64475	0,64788	0,6561	0,6215	0,6099	0,6206	0,3620	0,6561
Min	0,3019	0,3858	0,4978	0,5466	0,5946	0,6210	0,6303	0,6351	0,6293	0,6370	0,6215	0,6061	0,6153	0,3019	0,6399
Max	0,3620	0,4659	0,5411	0,5791	0,6331	0,6526	0,6544	0,6658	0,6579	0,6615	0,6518	0,6408	0,6433	0,3620	0,6658
Diff	0,06013	0,08013	0,04337	0,0325	0,0385	0,03163	0,02412	0,03063	0,02862	0,0245	0,0303	0,0346	0,028	0,06013	0,0259
P<0,05	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein
P<0,1	nein	nein	nein	nein	nein	ja	nein	ja	nein	ja	nein	ja	ja	nein	nein

Abbildung 5.18.: Klassifikationsergebnisse Binomial Naive Bayes, tabellarisch (Erstplatzierungen markiert)

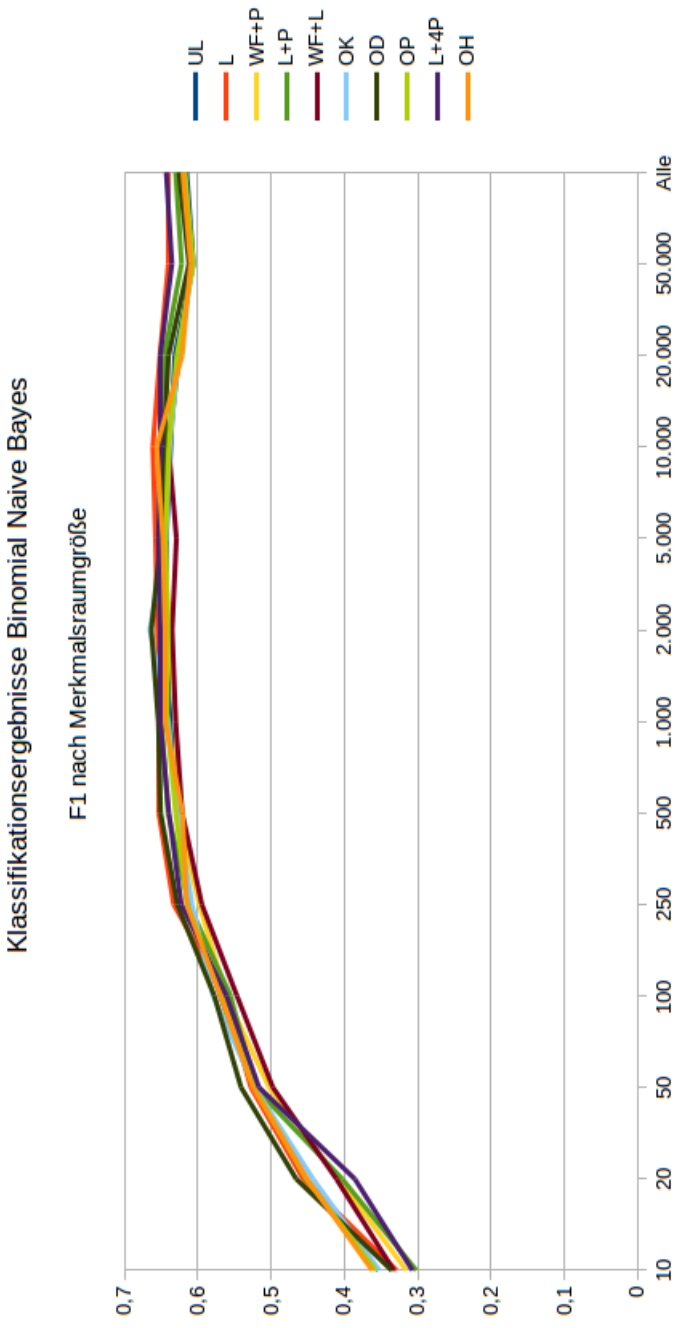
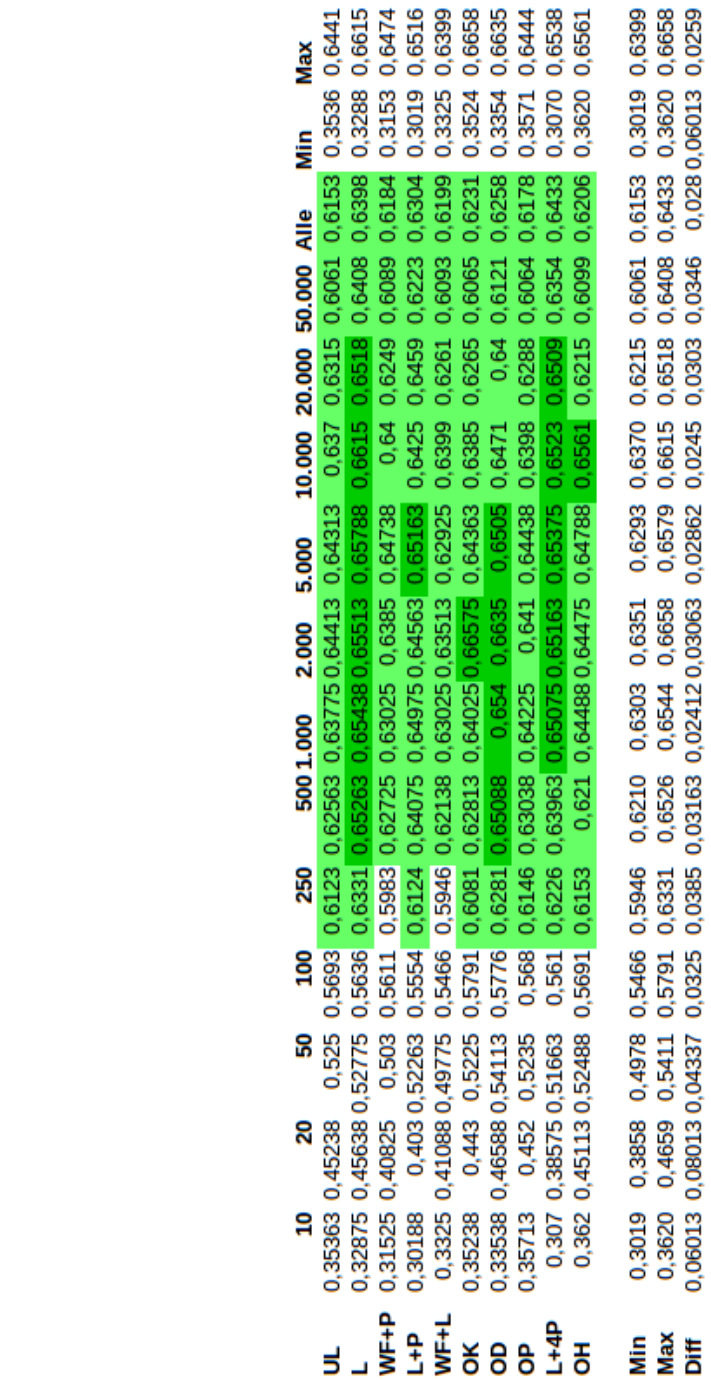


Abbildung 5.19.: Klassifikationsergebnisse Binomial Naive Bayes - F1 nach Merkmalsraumgröße

Auffälligerweise erreicht die Korpusversion *Ohne Konjugation* das beste Gesamtergebnis, wenn auch mit 2.000 Merkmalen bereits in einer der mittleren Merkmalsraumgrößen. Erneut können das Originalkorpus, die Versionen *Wortform+STTS* und *Wortform+Lemma* sowie zusätzlich die Modifikationen ohne abtrennbare Verbpartikeln keine Erstplatzierungen erreichen. In den kleineren Merkmalsräumen bis 100 Merkmale sowie erneut bei 2.000 Merkmalen dominieren die deklinations- und konjugationsbefreiten Korpusversionen, ehe die Führung von den beide Aspekte vereinenden lemmatabasierten Modifikationen übernommen wird. Zu beachten ist, dass nach acht Kreuzvalidierungen kein Ergebnis einer Korpusmodifikation einen statistisch als stark signifikant aufzufassenden P-Wert erreicht und mit fünf Ergebnissen nur eine Minderheit der Modifikationen schwach signifikante Verbesserungen gegenüber dem Originalkorpus erreicht.



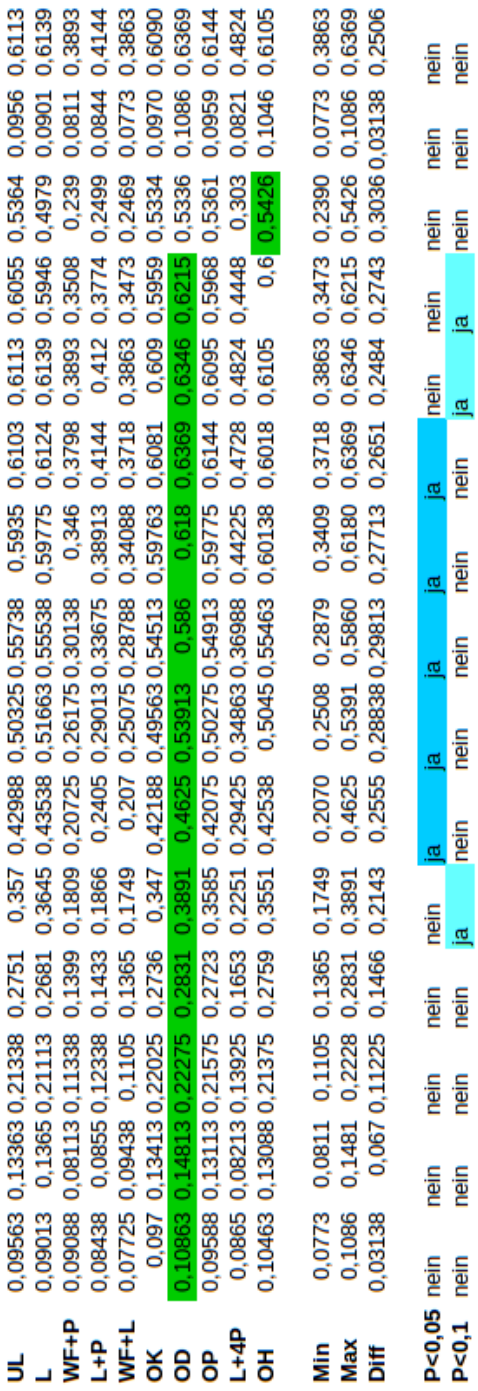
	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,35363	0,45238	0,525	0,5693	0,6123	0,62563	0,63775	0,64413	0,64313	0,637	0,6315	0,6061	0,6153	0,3536	0,6441
L	0,32875	0,45638	0,52775	0,5636	0,6331	0,65263	0,65438	0,65513	0,65788	0,6615	0,6518	0,6408	0,6398	0,3288	0,6615
WF+P	0,31525	0,40825	0,503	0,5611	0,5983	0,62725	0,63025	0,6385	0,64738	0,64	0,6249	0,6089	0,6184	0,3153	0,6474
L+P	0,30188	0,403	0,52263	0,5554	0,6124	0,64075	0,64975	0,64563	0,65163	0,6425	0,6459	0,6223	0,6304	0,3019	0,6516
WF+L	0,3325	0,41088	0,49775	0,5466	0,5946	0,62138	0,63025	0,63513	0,62925	0,6399	0,6261	0,6093	0,6199	0,3325	0,6399
OK	0,35238	0,443	0,5225	0,5791	0,6081	0,62813	0,64025	0,66575	0,64363	0,6385	0,6265	0,6065	0,6231	0,3524	0,6658
OD	0,33538	0,46588	0,54113	0,5776	0,6281	0,65088	0,654	0,6635	0,6505	0,6471	0,64	0,6121	0,6258	0,3354	0,6635
OP	0,35713	0,452	0,5235	0,568	0,6146	0,63038	0,64225	0,641	0,64438	0,6398	0,6288	0,6064	0,6178	0,3571	0,6444
L+4P	0,307	0,38575	0,51663	0,561	0,6226	0,63963	0,65075	0,65163	0,65375	0,6523	0,6509	0,6354	0,6433	0,3070	0,6538
OH	0,362	0,45113	0,52488	0,5691	0,6153	0,621	0,64488	0,64475	0,64788	0,6561	0,6215	0,6099	0,6206	0,3620	0,6561
Min	0,3019	0,3858	0,4978	0,5466	0,5946	0,6210	0,6303	0,6351	0,6293	0,6370	0,6215	0,6061	0,6153	0,3019	0,6399
Max	0,3620	0,4659	0,5411	0,5791	0,6331	0,6526	0,6544	0,6658	0,6579	0,6615	0,6518	0,6408	0,6433	0,3620	0,6658
Diff	0,06013	0,08013	0,04337	0,0325	0,0385	0,03163	0,02412	0,03063	0,02862	0,0245	0,0303	0,0346	0,028	0,06013	0,0259

Abbildung 5.20.: Klassifikationsergebnisse Binomial Naive Bayes, tabellarisch (F-Scores > 0,60 markiert)

Obgleich dieser Klassifikator mit 18 Ergebnissen mit einem F-Score oberhalb von 0,65 im Vergleich zu 46 Ergebnissen mit diesem Wert für die Logistische Regression bescheidener abschneidet, zeigt sich ein analoges Muster bei der Entwicklung in Richtung dieses Schwellenwertes im Hinblick auf die Merkmalsraumgröße: Wie bei der Logistischen Regression setzt eine deutliche Steigung über den Schwellenwert eines F1 von mehr als 0,60 überwiegend bei 250 Merkmalen und ausnahmslos bei 500 Merkmalen ein. Der bereits angesprochene Interpretationsspielraum bei der Evaluation der Klassifikatoren zeigt sich im Vergleich dieser beiden Klassifikatoren nach absolutem F1-Wert erneut: Unter Hinzunahme der Zellen über 0,60 (zusätzliche 70 für Binomial Naive Bayes sowie 41 für die Logistische Regression) ergibt sich ein minimaler Vorsprung von 88 zu 87 F-Scores oberhalb von 0,60 zugunsten des Bayes-Klassifikators.

5.2.2.4. Multinomial Naive Bayes

Der unter den konventionellen Klassifikatoren viertplatzierte und damit schwächste Klassifikator der genauer evaluierten Vierergruppe, *Multinomial Naive Bayes*, zeigt das kohärenteste Muster in Hinsicht auf Verbesserungen des Klassifikationsergebnisses durch flexionsbeschränkende Korpusmodifikationen, präsentiert in den Abbildungen 5.21 bis 5.23. Sämtliche bis auf eine der Erstplatzierungen inklusive des Gesamtbestwerts werden von der deklinationsbereinigten Korpusversion belegt. Die verbleibende Erstplatzierung erfolgt durch die homographiebefreite Korpusversion. Über die absolute Dominanz der Erstplatzierungen hinaus bemerkenswert ist des Weiteren, dass als einziger Klassifikator im Szenario Multinomial Naive Bayes bereits von Anfang an, also ab dem kleinsten Merkmalsraum von lediglich zehn Merkmalen, von einer Korpusmodifikation, hier der Entfernung der Deklination, profitiert.



	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,09563	0,13363	0,21338	0,2751	0,357	0,42988	0,50325	0,55738	0,5935	0,6103	0,6113	0,6055	0,5364	0,0956	0,6113
L	0,09013	0,1365	0,21113	0,2681	0,3645	0,43538	0,51663	0,55538	0,59775	0,6124	0,6139	0,5946	0,4979	0,0901	0,6139
WF+P	0,09088	0,08113	0,11338	0,1399	0,1809	0,20725	0,26175	0,30138	0,346	0,3798	0,3893	0,3508	0,239	0,0811	0,3893
L+P	0,08438	0,0855	0,12338	0,1433	0,1866	0,2405	0,29013	0,33675	0,38913	0,4144	0,412	0,3774	0,2499	0,0844	0,4144
WF+L	0,07725	0,09438	0,1105	0,1365	0,1749	0,207	0,25075	0,28788	0,34088	0,3718	0,3863	0,3473	0,2469	0,0773	0,3863
OK	0,097	0,13413	0,22025	0,2736	0,347	0,42188	0,49563	0,54513	0,59763	0,6081	0,609	0,5959	0,5334	0,0970	0,6090
OD	0,10863	0,14813	0,22275	0,2831	0,3891	0,4625	0,53913	0,586	0,618	0,6369	0,6346	0,6215	0,5336	0,1086	0,6369
OP	0,09588	0,13113	0,21575	0,2723	0,3585	0,42075	0,50275	0,54913	0,59775	0,6144	0,6095	0,5968	0,5361	0,0959	0,6144
L+4P	0,0865	0,08213	0,13925	0,1653	0,2251	0,29425	0,34863	0,36988	0,44225	0,4728	0,4824	0,4448	0,303	0,0821	0,4824
OH	0,10463	0,13088	0,21375	0,2759	0,3551	0,42538	0,5045	0,55463	0,60138	0,6018	0,6105	0,6	0,5426	0,1046	0,6105
Min	0,0773	0,0811	0,1105	0,1365	0,1749	0,2070	0,2508	0,2879	0,3409	0,3718	0,3863	0,3473	0,2390	0,0773	0,3863
Max	0,1086	0,1481	0,2228	0,2831	0,3891	0,4625	0,5391	0,5860	0,6180	0,6369	0,6346	0,6215	0,5426	0,1086	0,6369
Diff	0,03138	0,067	0,11225	0,1466	0,2143	0,2555	0,28838	0,29813	0,27713	0,2651	0,2484	0,2743	0,3036	0,03138	0,2506
P<0,05	nein	nein	nein	nein	nein	ja	ja	ja	ja	ja	nein	nein	nein	nein	nein
P<0,1	nein	nein	nein	nein	ja	nein	nein	nein	nein	nein	ja	ja	nein	nein	nein

Abbildung 5.21.: Klassifikationsergebnisse Multinomial Naive Bayes, tabellarisch (Erstplatzierungen markiert)

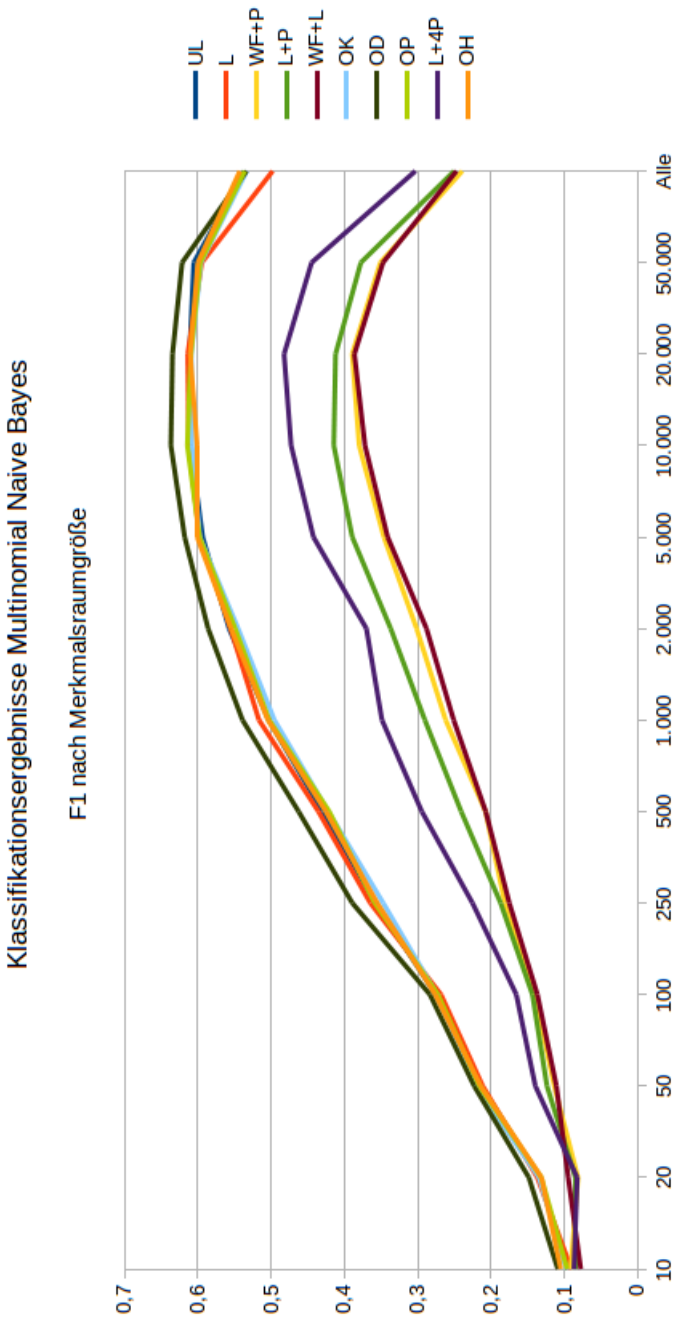


Abbildung 5.22.: Klassifikationsergebnisse Multinomial Naive Bayes - F1 nach Merkmalsraumgröße

Der Klassifikator erreicht sein bestes Gesamtergebnis bei 20.000 Merkmalen und verliert sodann zuerst gut einen und bei sämtlichen Merkmalen deutliche 7,89 Punkte beim F-Score. Unter Verzicht auf eine qualitative Analyse erscheint die Vermutung, dass eine Verdopplung bis Verdreifachung (vergleiche Tabelle 5.2 zur Anzahl Types) der zur Verfügung gestellten Merkmale angesichts der Quote niedrigfrequenter Merkmale mit hohem Artefaktanteil dem Klassifikator die statistische Einordnung dieser Merkmale erschwert, naheliegend. Eine derart varianzbehaftete Frequenzinformation könnte für das Mehrfachvorkommen explizit wertende multinomiale Modell des Naive Bayes-Klassifikators problematisch zu interpretieren sein und den deutlichen Bruch in der Entwicklung der F-Scores begründen. Bemerkenswerterweise weist dieser Klassifikator nach P-Werten (mit fünf mal $p < 0,05$ und weiteren drei P-Werten $< 0,10$) die statistisch zweitstärkste Absicherung der Verbesserungen durch Korpusmodifikationen inklusive seines besten Ergebnisses auf.

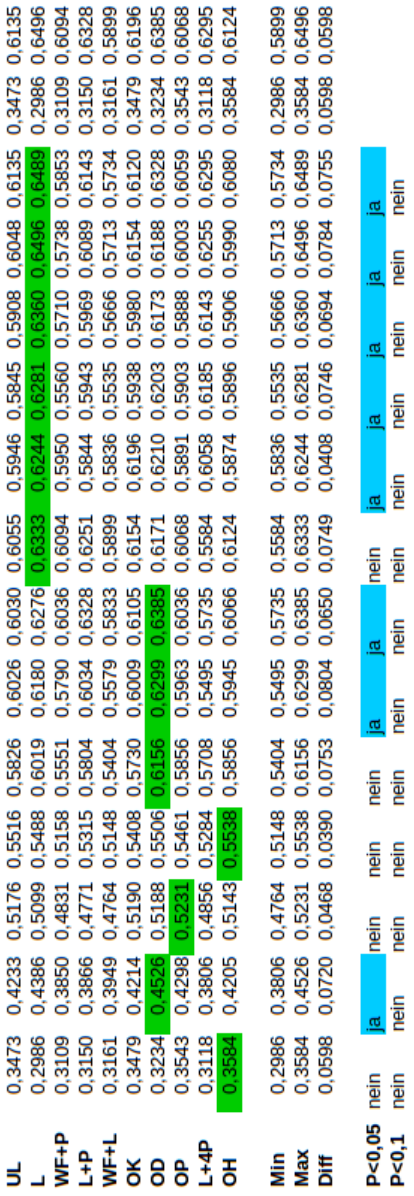
Die Viertplatzierung des Klassifikators nach Gesamtergebnissen spiegelt sich in ingesamt schwächeren absoluten Ergebnissen wider. F-Scores über 0,65 werden nicht erreicht, auch die Anzahl der Erstplatzierunge über 0,60 ist mit 17 bescheiden. Letzere sind überhaupt erst in der Merkmalsraumgröße 5.000 Merkmale vereinzelt zu finden und dominieren lediglich in den folgenden Größen 10.000 und 20.000 Merkmale. Der Verteilung der Erstplatzierungen folgend, dominiert auch hier die deklinationsfreie Korpusversion.

	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,09563	0,13363	0,21338	0,2751	0,357	0,42988	0,50325	0,55738	0,5935	0,6103	0,6113	0,6055	0,5364	0,0956	0,6113
L	0,09013	0,1365	0,21113	0,2681	0,3645	0,43538	0,51663	0,55538	0,59775	0,6124	0,6139	0,5946	0,4979	0,0901	0,6139
WF+P	0,09088	0,08113	0,11338	0,1399	0,1809	0,20725	0,26175	0,30138	0,346	0,3798	0,3893	0,3508	0,239	0,0811	0,3893
L+P	0,08438	0,0855	0,12338	0,1433	0,1866	0,2405	0,29013	0,33675	0,38913	0,4144	0,412	0,3774	0,2499	0,0844	0,4144
WF+L	0,07725	0,09438	0,1105	0,1365	0,1749	0,207	0,25075	0,28788	0,34088	0,3718	0,3863	0,3473	0,2469	0,0773	0,3863
OK	0,097	0,13413	0,22025	0,2736	0,347	0,42188	0,49563	0,54513	0,59763	0,6081	0,609	0,5959	0,5334	0,0970	0,6090
OD	0,10863	0,14813	0,22275	0,2831	0,3891	0,4625	0,53913	0,586	0,618	0,6369	0,6346	0,6215	0,5336	0,1086	0,6369
OP	0,09588	0,13113	0,21575	0,2723	0,3585	0,42075	0,50275	0,54913	0,59775	0,6144	0,6095	0,5968	0,5361	0,0959	0,6144
L+4P	0,0865	0,08213	0,13925	0,1653	0,2251	0,29425	0,34863	0,36988	0,44225	0,4728	0,4824	0,4448	0,303	0,0821	0,4824
OH	0,10463	0,13088	0,21375	0,2759	0,3551	0,42538	0,5045	0,55463	0,60138	0,6018	0,6105	0,6	0,5426	0,1046	0,6105
Min	0,0773	0,0811	0,1105	0,1365	0,1749	0,2070	0,2508	0,2879	0,3409	0,3718	0,3863	0,3473	0,2390	0,0773	0,3863
Max	0,1086	0,1481	0,2228	0,2831	0,3891	0,4625	0,5391	0,5860	0,6180	0,6369	0,6346	0,6215	0,5426	0,1086	0,6369
Diff	0,03138	0,067	0,11225	0,1466	0,2143	0,2555	0,28838	0,29813	0,27713	0,2651	0,2484	0,2743	0,3036	0,03138	0,2506

Abbildung 5.23.: Klassifikationsergebnisse Multinomial Naive Bayes, tabellarisch (F-Scores $> 0{,}60$ markiert)

5.2.2.5. Support Vector Machine

Die Support Vector Machine mit RBF-Kernel erreicht den vorletzten Platz der sechs getesteten konventionellen Klassifikatoren mit einem F-Score von 0,6496 unter Verwendung des Korpus *Lemmatisiert*. Diese Korpusversion stellt auch ausnahmslos die Erstplatzierungen der größeren Merkmalsräume ab 2.000 Merkmalen, die im T-Test gegenüber dem Originalkorpus durchgängig P-Werte $< 0,05$ aufweisen. In den kleineren Merkmalsräumen stellt die Korpusversion *Ohne Deklination* vier von sieben Erstplatzierungen, von denen drei als einzige dieser Merkmalsräume statistisch mit $p < 0,05$ gestützt werden. Mit insgesamt acht mit diesem Wert getesteten Verbesserungen erscheinen die SVM-Experimente auch insgesamt als statistisch überdurchschnittlich aussagekräftig. Die Ergebnisverteilung nach absoluten F-Score-Werten platziert die Support Vector Machine mit 50 Werten oberhalb von 0,60, darunter keine oberhalb von 0,65, lediglich auf Platz drei. Das Verteilungmuster entspricht dabei wiederum, wenn auch weniger konsequent und lückenhafter ausgeprägt, dem der Logistischen Regression und des binomialen Naive Bayes-Modells, mit erstmaligem Erscheinen bei 250 und prominenterer Ausprägung ab 500 Merkmalen. Den Ergebnissen der Erstplatzierungen folgend, dominieren auch hier die lemmatisierte und die deklinationsfreie Korpusversion.



	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,3473	0,4233	0,5176	0,5516	0,5826	0,6026	0,6030	0,6055	0,5946	0,5845	0,5908	0,6048	0,6135	0,3473	0,6135
L	0,2986	0,4386	0,5099	0,5488	0,6019	0,6180	0,6276	0,6333	0,6244	0,6281	0,6360	0,6496	0,6489	0,2986	0,6496
WF+P	0,3109	0,3850	0,4831	0,5158	0,5551	0,5790	0,6036	0,6094	0,5950	0,5560	0,5710	0,5738	0,5853	0,3109	0,6094
L+P	0,3150	0,3866	0,4771	0,5315	0,5804	0,6034	0,6328	0,6251	0,5844	0,5943	0,5969	0,6089	0,6143	0,3150	0,6328
WF+L	0,3161	0,3949	0,4764	0,5148	0,5404	0,5579	0,5833	0,5899	0,5836	0,5535	0,5666	0,5713	0,5734	0,3161	0,5899
OK	0,3479	0,4214	0,5190	0,5408	0,5730	0,6009	0,6105	0,6154	0,6196	0,5938	0,5980	0,6154	0,6120	0,3479	0,6196
OD	0,3234	0,4526	0,5188	0,5506	0,6156	0,6299	0,6385	0,6171	0,6210	0,6203	0,6173	0,6188	0,6328	0,3234	0,6385
OP	0,3543	0,4298	0,5231	0,5461	0,5856	0,5963	0,6036	0,6068	0,5891	0,5903	0,5888	0,6003	0,6059	0,3543	0,6068
L+4P	0,3118	0,3806	0,4856	0,5284	0,5708	0,5495	0,5735	0,5584	0,6058	0,6185	0,6143	0,6255	0,6295	0,3118	0,6295
OH	0,3584	0,4205	0,5143	0,5538	0,5856	0,5945	0,6066	0,6124	0,5874	0,5896	0,5906	0,5990	0,6080	0,3584	0,6124
Min	0,2986	0,3806	0,4764	0,5148	0,5404	0,5495	0,5735	0,5584	0,5836	0,5535	0,5666	0,5713	0,5734	0,2986	0,5899
Max	0,3584	0,4526	0,5231	0,5538	0,6156	0,6299	0,6385	0,6333	0,6244	0,6281	0,6360	0,6496	0,6489	0,3584	0,6496
Diff	0,0598	0,0720	0,0468	0,0390	0,0753	0,0804	0,0650	0,0749	0,0408	0,0746	0,0694	0,0784	0,0755	0,0598	0,0598
P<0,05	nein	ja	nein	nein	nein	ja	ja	nein	ja	ja	ja	ja	ja		
P<0,1	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein	nein		

Abbildung 5.24.: Klassifikationsergebnisse Support Vector Machine, tabellarisch (Erstplatzierungen markiert)

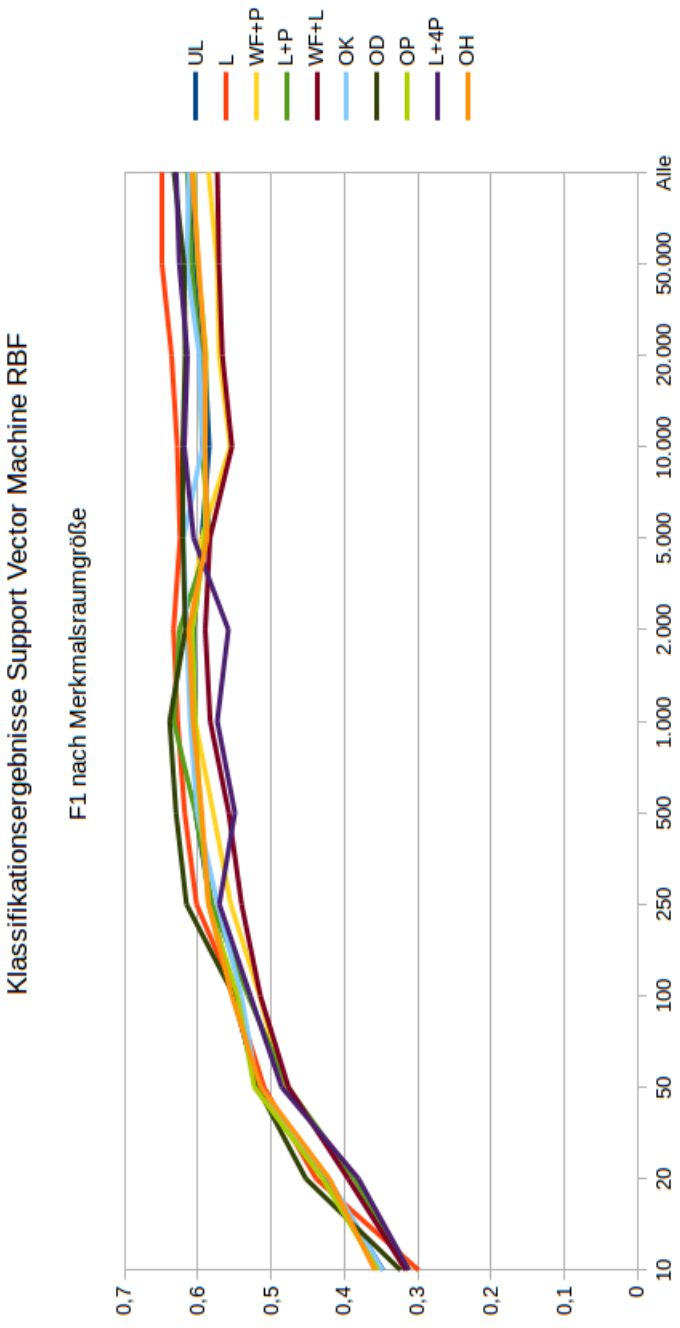


Abbildung 5.25.: Klassifikationsergebnisse Support Vector Machine - F1 nach Merkmalsraumgröße

	10	20	50	100	250	500	1.000	2.000	5.000	10.000	20.000	50.000	Alle	Min	Max
UL	0,34725	0,42325	0,51763	0,5516	0,5826	0,60263	0,603	0,6055	0,59463	0,5845	0,5908	0,6048	0,6135	0,3473	0,6135
L	0,29863	0,43863	0,50988	0,5488	0,6019	0,618	0,62763	0,63325	0,62438	0,6281	0,636	0,6496	0,6489	0,2986	0,6496
WF+P	0,31088	0,385	0,48313	0,5158	0,5551	0,579	0,60363	0,60938	0,595	0,556	0,571	0,5738	0,5853	0,3109	0,6094
L+P	0,315	0,38663	0,47713	0,5315	0,5804	0,60338	0,63275	0,62513	0,58438	0,5943	0,5969	0,6089	0,6143	0,3150	0,6328
WF+L	0,31613	0,39488	0,47638	0,5148	0,5404	0,55788	0,58325	0,58988	0,58363	0,5535	0,5666	0,5713	0,5734	0,3161	0,5899
OK	0,34788	0,42138	0,519	0,5408	0,573	0,60088	0,6105	0,61538	0,61963	0,5938	0,598	0,6154	0,612	0,3479	0,6196
OD	0,32338	0,45263	0,51875	0,5506	0,6156	0,62988	0,6385	0,61713	0,621	0,6203	0,6173	0,6188	0,6328	0,3234	0,6385
OP	0,35425	0,42975	0,52313	0,5461	0,5856	0,59625	0,60363	0,60675	0,58913	0,5903	0,5888	0,6003	0,6059	0,3543	0,6068
L+4P	0,31175	0,38063	0,48563	0,5284	0,5708	0,5495	0,5735	0,55838	0,60575	0,6185	0,6143	0,6255	0,6295	0,3118	0,6295
OH	0,35838	0,4205	0,51425	0,5538	0,5856	0,5945	0,60663	0,61238	0,58738	0,5896	0,5906	0,599	0,608	0,3584	0,6124
Min	0,2986	0,3806	0,4764	0,5148	0,5404	0,5495	0,5735	0,5584	0,5836	0,5535	0,5666	0,5713	0,5734	0,2986	0,5899
Max	0,3584	0,4526	0,5231	0,5538	0,6156	0,6299	0,6385	0,6333	0,6244	0,6281	0,6360	0,6496	0,6489	0,3584	0,6496
Diff	0,05975	0,072	0,04675	0,039	0,0753	0,08038	0,065	0,07488	0,04075	0,0746	0,0694	0,0784	0,0755	0,05975	0,0598

Abbildung 5.26.: Klassifikationsergebnisse Support Vector Machine, tabellarisch (F-Scores $> 0{,}60$ markiert)

5.2.2.6. Gesamtanalyse

Von größerer Relevanz als das absolute Abschneiden einzelner Klassifikatoren auf dem Untersuchungskorpus im Allgemeinen ist für den Zweck dieser Untersuchung eine algorithmenunabhängige vergleichende Analyse der Ergebnisse im Hinblick auf die Resultate der Korpusmodifikationen. Tabelle 5.9 zeigt zunächst den Anteil der einzelnen Korpusmodifikationen an den Erstplatzierungen bei den Klassifikatoren der Vierergruppe:

Korpus	**Gesamt**	**p<0,05**	**p<0,1**	**p mind. <0,1**
Unlemmatisiert	0	0	0	0
Lemmatisiert	17	5	6	11
Wortform+STTS	0	0	0	0
Lemma+STTS	2	2	0	2
Wortform+Lemma	0	0	0	0
Ohne Konjugation	3	0	1	1
Ohne Deklination	21	10	3	13
Ohne PTKVZ	2	0	0	0
Lemma+4POS	3	0	2	2
Ohne Homographen	4	0	0	0
Summe	52	17	12	29

Tabelle 5.9.: Erstplatzierungen Korpusversionen, klassifikatorübergreifend

Deutlich erkennbar dominiert die deklinationsfreie Korpusversion als einzelne Modifikation sowohl die Erstplatzierungen insgesamt als auch die Listen der beiden Signifikanzniveaus: Eine relative Mehrheit von 21 Erstplatzierungen insgesamt sowie von 10 Erstplatzierungen mit einem Signifikanzwert von $p < 0,05$ wird bei der Entfernung der Deklination durch Teillemmatisierung bei Substantiven und Adjektiven erreicht. Deklinationsfreie Korpusversionen belegen als einzige der neun Modifikationen in jedem Klassifikator der Vierergruppe mindestens einen ersten Platz und erreichen stets als erste oder gleichauf mit anderen Korpusversionen F-Scores von über $0,6$. Der Deklination als einzeln benennbarem Flexionsphänomen folgt auf dem zweiten Platz die Lemmatisierung als methodische Korpusmodifikation mit 17 Erstplatzierungen, darunter ebenfalls zweitstärksten 5

Erstplatzierungen mit Signifikanzniveau $p < 0,05$. Sie erreicht F-Scores von über $0,60$ im Wesentlichen zeitgleich mit deklinationsfreien Korpora und außer in der Support Vector Machine häufiger F-Scores über $0,65$. Den dritten Platz belegt die ebenfalls lemmabasierte Korpusversion *Lemma+STTS* mit zwei Erstplatzierungen (mit $p < 0,05$). Es folgen die konjugationsfreie Korpusversion und die lemmatisierte Korpusversion mit kleinem POS-Tagset mit jeweils drei Erstplatzierungen (davon zwei respektive eine mit $p < 0,1$). Keine statistisch signifikanten Erstplatzierungen die homographiefreie und die Korpusversion ohne abtrennbare Verbpartikeln mit vier respektive zwei Erstplatzierungen. Keinerlei Erstplatzierungen erringen das Originalkorpus sowie die Modifikationen *Wortform+STTS* und *Wortform+Lemma*. Die drei erfolgreichen lemmatabasierten Korpusmodifikationen gemeinsam gewertet überholen die deklinationsfreie Version bei der Gesamtzahl der Erstplatzierungen mit nunmehr 22 zu 21, nicht jedoch bei den Erstplatzierungen mit $p < 0,05$.

	10	20	50	100	250	500	1.000	2k	5k	10k	20k	50k	Alle
UL													
L				1	1	2	1	2	2	2	2	3	1
WF+P													
L+P									1	1			
WF+L													
OK	1			1				1					
OD	1	4	2	1	3	2	3	1	1	1	1	1	
OP			2										
L+4P											1		2
OH	2			1									1

Tabelle 5.10.: Erstplatzierungen Korpusversionen nach Merkmalsraumgrößen

Tabelle 5.10 und Abbildung 5.27 wechseln die Perspektive auf die Größe der Merkmalsräume. Um den vorhandenen Trend klar darstellen zu können, werden an dieser Stelle die Erstplatzierungen unabhängig ihrer Signifikanzniveaus verwendet. Die Übersichten zeigen überwiegende Erstplatzierungen der deklinationsfreien Korpusversion in den kleinen bis mittleren Merkmalsräumen bis 1.000 Merkmale. Sodann erfolgt die Ablösung

durch die bereits seit Merkmalsraumgröße 100 auftretende lemmatisierte Korpusversion, die sämtliche weiteren Merkmalsräume bis 50.000 Merkmale dominiert. Die Korpusversion *Lemmatisiert* verliert diese Führung im Fall sämtlicher verfügbarer Merkmale an ihre Erweiterungsversion mit kleinem POS-Tagset. Sie teilt sich nunmehr den zweiten Platz mit der wieder erscheinenden homographiebefreiten Korpusversion. Die konjugationsbezogenen Korpora *Ohne Konjugation* und *Ohne PTKVZ* schneiden mit gemeinsam gewerteten fünf von 52 Erstplatzierungen schwach ab. Sie sind speziell in kleineren Merkmalsräumen bis 100 Merkmale anzutreffen, mit der beim binomialen Bayes-Klassifikator beobachteten Ausnahme für *Ohne Konjugation* bei 2.000 Merkmalen. Bemerkenswert ist das Erstplatzierungsmuster der homographiefreien Korpusmodifikation: Bei insgesamt eher schwachem Abschneiden und Auftreten zunächst nur in sehr kleinen Merkmalsräumen erscheint sie erstplatziert erneut bei Verwendung sämtlicher Merkmale (im multinomialen Bayes-Klassifikator). Ihr dortiges gemeinsames Auftreten gemeinsam mit den Versionen *Lemmatisiert* und *Lemma+4POS* deutet darauf hin, dass in dieser maximalen Merkmalslistengröße die Disambiguierung von Homographen eine gewachsene Rolle bei der Besetzung der zahlreichen zusätzlichen Listenplätze spielt. Diese hat die nicht flexionsbereinigte, aus Wortformen und dem verkleinerten POS-Tagset gebildete Korpusversion mit den beiden erstgenannten gemeinsam. Keinerlei Erstplatzierungen belegen wie bereits bekannt das Originalkorpus sowie die beiden wortformenbasierten Versionen *Wortform+STTS* sowie *Wortform+Lemma*. Die beiden letztgenannten modifizierten Korpusversionen verfügen ausweislich Tabelle 5.5 in Abschnitt 5.2.1 über die meisten Types aller Korpusversionen. Die erhebliche Vergrößerung der Typesmenge gegenüber dem in dieser Hinsicht viertplatzierten Originalkorpus geht nachvollziehbarer Weise nicht mit einem hinreichenden Informationsgewinn für die Klassifikatoren einher.

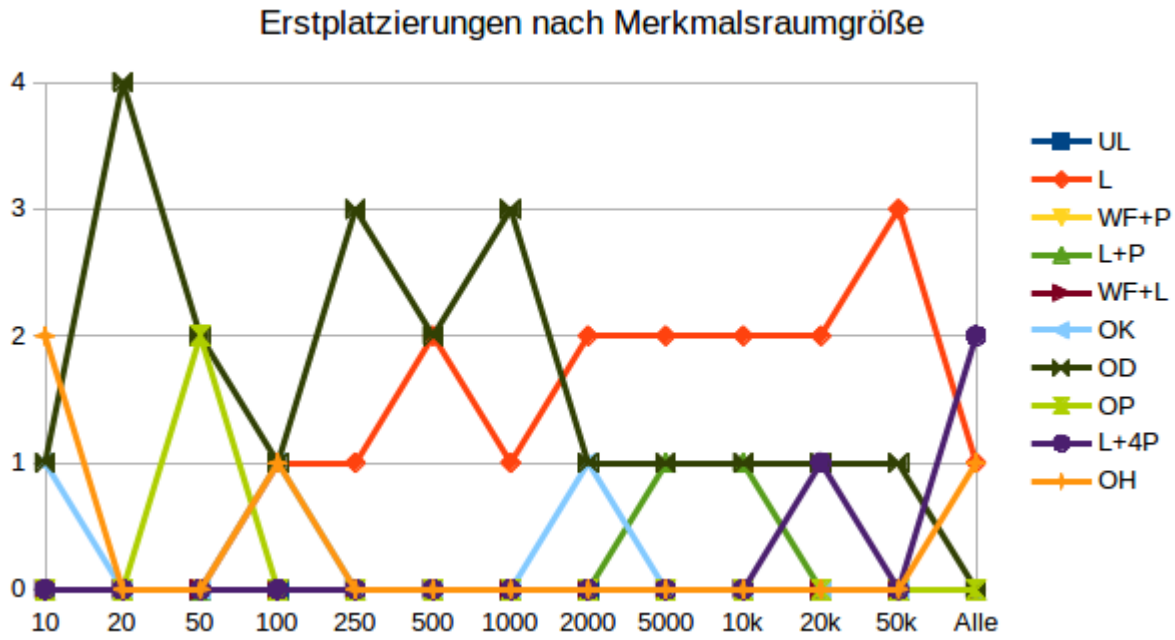


Abbildung 5.27.: Erstplatzierung Korpusversionen nach Merkmalsraumgrößen

5.2.3. Zusammenfassung und Diskussion konventionelle Klassifikation

Die vorstehende vergleichende Gesamtanalyse der konventionellen Klassifikationsergebnisse weist zunächst die Deklination 21 Erstplatzierungen als das mit großem Abstand relevanteste Flexionsphänomen im Prozess der konventionellen Klassifikation im vorliegenden Szenario aus. Angesichts des geringen Anteils von lediglich gut 10% der Erstplatzierungen (5 von 52 Gesamtergebnissen, darunter keines mit einem Signifikanzniveau von $p < 0,05$) der konjugationsbezogenen Korpusversionen ist möglicherweise sogar die Aussage angemessen, bei der Deklination handle es sich um die einzige im konventionellen Klassifikationsgeschehen überhaupt relevante Flexionserscheinung. Dieses eindeutige Ergebnis bestätigt wesentliche Beobachtungen aus Kapitel 4: Die Unterkapitel 4.1 und 4.2

zeigen eine ausgeglichenere Verteilung der Werte der Deklinationskategorien Kasus und Numerus bei den Substantiven und Adjektiven als der verbalen Kategorien (in den Unterkapiteln 4.3 und 4.4). Gleichzeitig handelt es sich bei der Aufteilung der Kategoriewerte bei den deklinierten Wortklassen keinesfalls um eine Gleichverteilung. Die Verteilung wird überdies in intransparenter Weise von den in den genannten Abschnitten besprochenen Synkretismen überlagert. In Verbindung mit der in der Wortklassenanalyse in Abschnitt 5.1.1 aufgezeigten Dominanz dieser Wortklassen, insbesondere der Substantive, in den Merkmalsräumen bestätigt sich das in Kapitel 4 vermutete Potenzial für Lernerschwernisse hier nun als empirisch relevant. Der hohe Anteil von Substantiven in der Mehrheit der Merkmalsräume kann semantische Ursachen haben (die in der Klasse kondensierten Informationen werden in diesen Substantiven manifestiert), aber auch dem in Abschnitt 4.6.1 diskutierten Effekten geringerer Formenvielfalt auf die χ^2-basierte Auswahl geschuldet sein. Die in den Experimenten in den Abschnitten 5.1.1 bis 5.1.3 andeutungsweise beobachteten Interaktionen zwischen Semantik und Flexionsmorphologie deuten darauf hin, dass diese Ursachen nicht weiter anteilig quantifiziert werden können.

Mit abnehmender Dominanz deklinierter Merkmale in den Merkmalsräumen schwindet konsequenterweise die Effektivität der Entfernung der Deklination. Verben mit ihrer größeren Formenvielfalt übernehmen zunehmend, wenn auch weniger stark als nach der ersten Korpusanalyse zu erwarten, die Besetzung zusätzlicher Plätze auf den Merkmalslisten. Parallel steigt nach dem Gesetz der großen Zahlen die Wahrscheinlichkeit zusätzlich auftretender, seltenerer deklinierter Formen bereits bekannter Lexeme. Das Auftreten dieser zusätzlichen flektierten Formen insgesamt wiederum erhöht die Anzahl auftretender Homographen durch die Verwendung homographieverursachender Morpheme wie etwa *-en* und *-er* (siehe Unterkapitel 4.5, insbesondere Tabelle 4.36f). Allen genannten Vorgängen wirkt eine Lemmatisierung stark entgegen; die zusätzliche Verwendung des kleinen POS-Tagsets zur Kennzeichnung der Zugehörigkeit zu einer der drei offenen

Wortklassen in der Modifikation *Lemma+4POS* eliminiert sämtliche Phänomene dann vollständig (siehe Tabelle 5.6). Die um das kleine POS-Tagset erweiterte lemmatisierte Korpusversion ist als einzige außer ihrem STTS-erweiterten Pendant *Lemma+STTS* hierzu in der Lage (ibd.), benötigt jedoch signifikant weniger Terminalsymbole zur Kodierung der gleichen Menge relevanter Informationen (s. Tabelle 5.5). Das bereits im vorstehenden Abschnitt angemerkte starke gemeinsame Abschneiden der Korpusversionen *Lemmatisiert*, *Lemma+4POS* und *Ohne Homographen*, wobei Letzere alternativ adäquat auch als *Wortform+4POS* bezeichnet werden könnte, zeigt, dass für die Nutzbarmachung der vollen potenziellen Merkmalsmenge diese Phänomene kontrolliert werden müssen.

Eine zentrale Schlussfolgerung aus der Gegenüberstellung der flexionsmorphologischen Produktivität der drei offenen Wortklassen laut Flexionsparadigmen, ihrer empirischen Realisierung laut Korpusanalyse und ihrer Einflüsse auf die konventionellen Klassifikationsprozesse ist, dass sprachliche Konvention (in Form von empirischer Distribution) theoretische Formenvielfalt schlägt. Die Korpusanalyse zeigt eindeutig, dass in den formalen Nachrichtentexten nur ein geringer Teil der möglichen Formen der starken wie schwachen Verben realisiert werden und die Substantive aufgrund ihrer empirischen Dominanz sowohl im Korpus als auch dem hieraus extrahierten Merkmalsraum den deutlich größeren, messbaren Einfluss auf die erreichbaren Klassifikationsergebnisse ausüben.

Eine weitere Erkenntnis der empirischen Analysen in Kapitel 4 war die Feststellung, dass abgetrennte Verbpartikeln und Homographen in an Ubiquität grenzender Verbreitung (s. Tabellen 4.31 und 4.38) in der überwiegenden Anzahl der Texte teilweise in großer Anzahl anzutreffen waren. Während die geringe Relevanz des ersten Phänomens im Klassifikationsprozess dem geringen Anteil der Verben an der Modellierung der Kategorien zugeschrieben werden kann, erscheint die Entwicklung der Homographen komplexer. Ihr moderates, aber sichtbares Auftreten in kleinen Merkmalsräumen besitzt möglicherweise Artefaktcharakter. Ihr direkt (durch die Korpusversion *Ohne Homographen*) und indirekt (durch die lemmatisierten Korpusversionen) sichtbarer großer Einfluss im größten

Merkmalsraum kann hingegen möglicherweise unter Berücksichtigung der Zahlen aus den Tabellen 4.33 und 4.35 interpretiert werden: Ein großer Teil der Homographien entfällt auf Hilfs- und Modalverben, die für die weite Verbreitung von Homographen auf Textebene sorgen; semantisch relevante homographe Merkmale treten offensichtlich erst auf, nachdem alternativ bereits zahlreiche zusätzliche Merkmalslistenplätze von spezialisierten Verben und Adjektiven belegt wurden.

Nachdem unter den flexionsmorphologischen Phänomenen nur die Deklination (mit dem eindeutigen Schwerpunkt Substantivdeklination) als relevant identifiziert werden kann, ist grundsätzlich festzustellen, dass Flexionsmorphologie allgemein im vorliegenden konventionellen Klassifikationsszenario eine geringe empirische Rolle spielt, wenn sie mit dem Einfluss der Parameter Trainingsmenge und Merkmalsraumgröße verglichen werden: Der erreichbare F-Score verbessert sich bei allen Klassifikatoren der Vierergruppe bis auf den multinomialen Bayes-Klassifikator mit zunehmender Merkmalsraumgröße bemerkenswert kongruent von einem F-Score von eingangs rund $0,3$ auf Werte um $0,65$ bis $0,67$; beim genannten Bayes-Modell liegt der merkmalsraumabhängige Zuwachs mit einem Anstieg von etwa $0,1$ sogar etwa beim Faktor sechs. Die Analysen zum Klassifikationserfolg unter Bezug auf die Größe der Trainingsmenge, in den Abbildungen 5.16 bis 5.17 für die Logistische Regression stellvertretend dargestellt, zeigt einen ähnliche starken Einfluss der Anzahl der vorgestellten Trainingsdokumente. Überdies ist aus dieser Analyse ersichtlich, dass für einen hohen Klassifikationserfolg das optimale Verhältnis zwischen Merkmalsraumgröße und Trainingsmenge entscheidend und um ein Vielfaches wichtiger ist als Flexionseigenschaften der Merkmale: Eine vergrößerte Trainingsmenge führt nur zu Ergebnisgewinnen, wenn die aus ihr erlernbaren Korrelationen zwischen Merkmalen und Klassen auch in einer hinreichend großen Merkmalsliste abgelegt werden können. Umgekehrt leidet ein Merkmalsraum bei inadäquater Größe der Trainingsmenge an Unterspezifikation – die zur Verfügung stehenden Plätze können nicht mit hilfreichen Informationen belegt werden und beginnen zu „rauschen“.

Möglicherweise spiegelt der bereits in Unterabschnitt 5.2.2.5 besprochene Leistungsabfall des multinomialen Bayes-Klassifikators bei Zuführung aller verfügbaren Merkmale diese Zusammenhänge: Das Modell ist offensichtlich nicht in der Lage, die zusätzlichen Informationen der mehr als verdoppelten Merkmalsmenge gewinnbringend zu nutzen. Kombiniert mit dem Umstand, dass nur bei diesem Klassifikator die homographiebereinigte Korpusversion in diesem Merkmalsraum eine Erstplatzierung erreicht, lässt sich möglicherweise schließen, dass ein relevanter Anteil selten vorkommender Merkmale, etwa flektierter Verbformen, mit Homographen belegt ist, die aufgrund der sowohl eigenen geringen Frequenz als auch der Seltenheit des Merkmals starke Verzerrungen der gemessenen Korrelation verursachen. Grundsätzlich ist also anzumerken, dass die Zusammenstellung großer Merkmalsräume aus angemessen großen Trainingskorpora den Klassifikationserfolg in den meisten Fällen verbessert. Bis zu welchem Ausmaß diese Verbesserung funktionieren kann und welche Aufbereitung die Merkmale bei der Übernahme aus dem unlemmatisierten Korpus erfahren müssen, um mehr Informationen als Rauschen in den Merkmalsraum einzubringen, ist jedoch offensichtlich klassifikatorspezifisch.

Die abschließende Schlussfolgerung der experimentellen Untersuchungen dieses Unterkapitels in gemeinsamer Interpretation mit der theoretischen und empirischen Analyse des Kapitels 4 und den Merkmalsraumanalyse des Unterkapitels 5.1 lautet daher: Konventionelle Klassifikation wird im vorliegenden exemplarischen Szenario nur zu einem geringen Anteil von flexionsmorphologischen Prozessen beeinflusst. Deren Relevanz erscheint gering im Vergleich zu Trainingskorpusgröße und Merkmalsraumgröße und deren Verhältnis zueinander. Nichtsdestotrotz ist dieser verhältnismäßig geringe Einfluss in 17 von 52 Ergebnissen deutlich signifikant, in einer absoluten Mehrheit von 29 Konstellationen schwach signifikant messbar. Die laufzeitbedingt lediglich moderate Anzahl von je acht Kreuvalidierungen der 130 Experimente verhindert allerdings möglicherweise eine statistisch höhere Aussagekraft der experimentellen Ergebnisse in Form zusätzlicher sichtbarer statistischer Signifikanz. Schließlich ermöglichte die unlemmatisierte Originalversion des

Korpus in keinem einzigen Fall ein besseres Klassifikationsergebnis als irgendeine der vorgenommenen Korpusmodifikationen. Selbst eine möglicherweise verborgene stärkere statistische Signifikanz dieser positiven Abstände ließe jedoch ohne Weiteres kaum erwarten, dass ihr relativer Einfluss gegenüber den Basisparametern der Klassifikatoren signifikant stärker ausfallen würde.

Kann nun die Flexionsmorphologie als in sichtbarem, aber geringem Ausmaß den Klassifikationsprozess eingeschätzt werden, gilt dies aus methodischer Sicht analog für die Lemmatisierung als Korpusmodifikationstechnik: Sämtliche Korpusmodifikationen beruhen letztlich auf entweder vollständiger Lemmatisierung oder einer wortklassenbezogenen Teillematisierung. Das homographieeleminierende Zusammensetzen eines Merkmals aus einer lemmatisierten Wortform und einem POS-Tag erweist sich als potenziell ewinnbringend, wenn es sich hierbei um ein nur die nötigsten Informationen (hier: Vier Tags zur Zugehörigkeit zu einer der offenen Wortklassen oder einer sonstigen Wortklasse) enthaltendes Tagset handelt.

Das Abschneiden der getaggten und expandierten Korpora demonstriert, dass nach menschlicher Lesart maximal informative, neu zusammengesetzte Merkmale, namentlich die der Korpusversionen *Wortart+STTS* sowie *Wortform+Lemma*, keinesfalls die optimale Informationsdichte für einen automatischen Klassifikationsprozess bereitstellen: Eine stark erhöhte Anzahl von unterschiedlichen Zeichenketten geht nicht automatisch mit einem proportionalen Informationsgewinn für den Klassifikator einher. Dieser ist bei der Interpretation der zur Verfügung gestellten Informationen auf ein ausgewogenes Verhältnis der Anzahl der Beobachtungen und ihrer Aussagekraft angewiesen, um bei maximalem Informationsgewinn einen Ausgleich zwischen Unterspezifikation und Overfitting seines Modells zu treffen.

5.3. Neuronale Klassifikation am Beispiel eines Convolutional Neural Network

Dieses Kapitel stellt die Experimente vor, die mit einem architektonisch moderat komplexen Convolutional Neural Network unter Verwendung von FastText-Embeddingvektoren auf dem Originalkorpus und den bereits eingeführten neun Modifikationen durchgeführt wurden. Das benötigte Embedding-Eingabeformat erforderte anstelle der bloßen Übersetzung der Wortformen im Text in ihre Embeddingvektoren für einige Korpusversionen eine separate Vorverarbeitung. Diese Verarbeitungsschritte werden im folgenden Abschnitt 5.3.1 beschrieben. Abschnitt 5.3.2 dokumentiert die Architektur des verwendeten Netzes nach einer Teilsuche in einem Raum möglicher Parameter. Abschnitt 5.3.3 präsentiert und analysiert die erzielten Ergebnisse in einem an die Beschreibung der konventionellen Klassifikatoren angelehnten Format.

5.3.1. Embeddingbasierte Korpusmodifikationen

Abweichend von der Gruppierung der Korpusversionen nach dem Grad linguistischer Bearbeitung in Unterabschnitt 5.2.1.1 lassen sich die Modifikationen des Originalkorpus für die Eingabe in das neuronale Netz in zwei Gruppen unterschiedlicher Vektorisierungsverfahren unterteilen. Die Korpusversionen *Unlemmatisiert*, *Lemmatisiert*, *Ohne Deklination*, *Ohne Konjugation* und *Ohne PTKVZ* basieren unabhängig von Komplett- oder Teillemmatisierung ausschließlich auf Wortformen ohne Ergänzungen auf Zeichenkettenebene. Die Umwandlung der Texte in Vektorform erfolgt somit durch einfachen Austausch jeder Wortform gegen ihren FastText-Vektor und die Konkatenierung dieser Embeddingvektoren in Eingabereihenfolge. Die für den Eingabelayer des neuronalen Netzes erforderliche einheitliche Länge der Textvektoren wurde durch Padding, also Auf-

füllen mit Nullstellen, auf eine Länge von 1.000 Tokens, sichergestellt, womit die Länge dieser Vektoren stets exakt 300.000 Stellen betrug.

Die Korpusversionen *Wortform+STTS*, *Lemma+STTS*, *Wortform+4POS* und *Lemma+4POS* ergänzen, wie in Unterabschnitt 5.2.1.1 dokumentiert, jede Wortform beziehungsweise jedes Lemma um einen POS-Tag aus dem STTS oder einem verkleinerten, vierteiligen Tagset. In diesen Modifikationen entsteht durch die Hinzufügung des POS-Tags formal gesehen ein neues Symbol durch eine neugeschaffene Zeichenkette. Deren TF-IDF kann ebenso ermittelt werden wie die einer zugrunde liegenden Wortform: Der Vektorisierer in der Vorverarbeitungsstufe von Scikit-learn besitzt keinerlei „Verständnis“ von Morphologie oder Semantik einer Zeichenkette. Der numerische Wert für den Eingabevektor wird lediglich durch das Auszählen und Normalisieren der Häufigkeit eines Symbols in Korpus und Einzeltext erzeugt.

Dieses Vorgehen kann beim Austausch derart modifizierter Wortformen gegen Embeddingvektoren nicht funktionieren: Während das Aufsuchen von „Häuser“ in der Embeddingtabelle den benötigten Vektor zutage fördert, scheitert diese Abfrage offensichtlich bei den Zeichenketten *Häuser+NN* (Korpusversion *Wortform+STTS*), *Haus+NN* (Korpusversion *Lemma+STTS*), *Häuser+S* (Korpusversion *Ohne Homographen*) und *Haus+S* (Korpusversion *Lemma+4POS*). Die Wortformen und Lemmata werden daher zunächst analog zu den oben stehenden Beispielen ohne Ergänzungen gegen ihre FastText-Vektoren ausgetauscht. Im Anschluss wird der Vektor um den POS-Tag in Form einer 301. Dimension als Ganzzahl ergänzt. Bei den STTS-Tags handelt es sich hierbei um eine Zahl von 1 bis 51, bei dem simplifizierten Tagset entsprechend um eine Zahl von 1 bis 4. Das Ergänzen der konkatenierten Vektoren um Padding-Nullstellen auf das Äquivalent von 1.000 Tokens erfolgt analog zu den übrigen Korpusversionen, so dass diese Korpusversionen eine fixe Eingabevektorlänge von stets 301.000 Stellen aufweisen.

Eine dritte Konvertierungsmethode generierte die verbleibende Korpusversion, *Wortform+Lemma*. In dieser Version wurden die Vektoren von Wortform und Lemma separat aus dem Embedding extrahiert und unmittelbar konkateniert. Die so entstehenden Vektoren von 600 Dimensionen Länge wurden mit denen der folgenden Paare analog zu den übrigen Korpusversionen aneinandergereiht. Der entstehende Textvektor wurde analog zu den bisherigen Korpusversionen auf die fixe Länge von 1.000 Tokens gepaddet. Ein Text wurde somit in dieser Version stets mit 600.000 Dimensionen modelliert.

Wie bereits in Unterkapitel 5.1 auf die Merkmalsraumebene bezogen thematisiert, sind nicht alle in TübaDZ vorkommenden flektierten Wortformen und noch weniger Lemmata in FastText enthalten. Zur Sicherstellung fixer Vektorenlänge wurden nicht aufzufindende Wortformen beziehungsweise Lemmata durch jeweils 300 beziehungsweise 301 Nullstellen ersetzt, die dann vom Klassifikator unberücksichtigt bleiben konnten. Der Abdeckungsgrad des Embeddings in Hinblick auf die in TübaDZ vorkommenden Types unterscheidet sich dabei wie in Tabelle 5.11 zu sehen erheblich zwischen Wortformen und Lemmata und zwischen den Wortklassen.

Klasse und Typ	**TübaDZ**	**FastText**	**Entspricht**
Adjektive dekliniert	27.292	22.387	82,03%
Adjektive Lemmata	15.349	10.739	69,97%
Substantive dekliniert	78.193	49.767	63,65%
Substantive Lemmata	65.222	38.559	59,12%
Verben konjugiert	16.980	15.180	89,40%
Verben Lemmata	7.611	3.478	45,70%

Tabelle 5.11.: Abdeckungsgrad Types TübaDZ in FastText

Der stark unterschiedliche Abdeckungsgrad verschiedener Wortklassen sowohl bei flektierten Formen als auch Lemmata ist unmittelbar ersichtlich, ebenso ein unterschiedlich starkes Abfallen der Abdeckung von flektierten zu lemmatisierten Types. Die Gründe für diesen unterschiedlichen Abdeckungsgrad wurden nicht weiter untersucht, sind aber möglicherweise aus dem Einfluss von Wikipedia als Teilgrundlage der Embedding-

Distribution zu sehen: Die formalen Texte der Enzyklopädie könnten eine ähnliche Distribution etwa der Verben aufweisen wie das Nachrichtentextkorpus und die Zitierform, äquivalent zur ersten und dritten Person Plural Indikativ Aktiv Präsens, seltener verwenden. Des Weiteren kann nicht vorausgesetzt werden, dass eine größere Anzahl domänenspezifischer Termini aus Nachrichtentexten der 1990er-Jahre in Gestalt der Substantive als häufigste Wortklasse in der inhaltlich breiter aufgestellten Wikipedia der 2000er und 2010er zu finden ist.

5.3.2. Architektur des künstlichen neuronalen Netzes

Der Beschränkung der Untersuchung auf unigrammbasierte Bag-of-Words-Modelle folgend, wurde ein Convolutional Neural Network als neuronaler Klassifikator gewählt. Auf kontextverarbeitende RNN-/LSTM-Komponenten wurde aus diesem Grund verzichtet. Zu bestimmende Parameter für den Aufbau des künstlichen neuronalen Netzes waren somit die Anzahl der Convolutional Layer, deren Filteranzahlen und Kernelgrößen, die Anzahl der Dense Layer und die Zahl ihrer Einheiten sowie die Aktivierungsfunktionen. Die Trainingsparameter Batch Size, Trainingsepochen und Validation Split ergänzen diese Konfigurationsmöglichkeiten für jedes erstellte Netz. Aufgrund der offensichtlich zahlreichen Parametrisierungsmöglichkeiten durch unterschiedliche Belegungen dieser Variablen und deren Kombination konnte in der zur Verfügung stehenden Zeit keine vollständige rasterartige Suche in diesem Parameterraum durchgeführt werden, so dass eine Suche entlang verschiedener Achsen erfolgte. Ein zumindest lokales Maximum konnte in der folgenden Konfiguration, die somit als Basis der weiteren Untersuchung diente, gefunden werden:

Auf den im vorhergehenden Abschnitt beschriebenen, aus den embeddeten und gepaddeten Texten bestehenden Eingabelayer folgen drei relu-aktivierte Convolutional Layer mit je 128 Filtern und einer Kernelgröße von 5, an die sich ein GlobalMaxPooling-Layer

anschließt. Die durch diese Kaskade extrahierten Merkmale wurden von einem ebenfalls relu-aktivierten Dense Layer mit 128 Einheiten verarbeitet, an den sich ein softmax-normalisierender Dense Layer mit acht die Zielklassen repräsentierenden Einheiten als Ausgabelayer anschließt.

5.3.3. Ergebnisse und Analyse der neuronalen Klassifikationsexperimente

Das Training des neuronalen Netzes in der ausgewählten Architektur erfolgte in Batchgrößen von je fünf Texten über 15 Epochen mit 10% der Texte als Validation-Split. Ein abschließender Test erfolgte auf stets 100 weiteren, unbekannten Texten. Zur Ermittlung des Lernverhaltens wurde jede Korpusversion auf 25 um je vier Prozent, entsprechend 136 Texten, wachsenden Korpusgrößen von 136 bis 3.400 Texten trainiert. Wie bei den konventionellen Klassifikatoren wurde auch hier die Trainingsmenge in jedem Wachstumsschritt vollständig neu zusammengestellt und nicht lediglich durch das Hinzufügen weiterer Texte vergrößert. Aufgrund großer Varianz der F-Scores erfolgte eine 40fache Kreuzvalidierung jedes Experiments. Tabelle 5.12 und Abbildung 5.28 zeigen den Lernverlauf des neuronalen Netzes mit wachsendem Trainingskorpus in der genannten Schrittweite.

T	UL	L	WF+S	L+S	WF+L	OD	OK	OP	WF+4	L+4
1	0,603	0,547	0,373	0,318	0,588	0,601	0,517	0,579	0,560	0,559
2	0,631	0,619	0,467	0,419	0,603	0,633	0,638	0,614	0,619	0,639
3	0,655	0,653	0,517	0,533	0,631	0,656	0,662	0,671	0,655	0,664
4	0,667	0,663	0,580	0,551	0,661	0,664	0,675	0,680	0,661	0,680
5	0,691	0,679	0,599	0,578	0,663	0,670	0,661	0,668	0,683	0,657
6	0,677	0,684	0,616	0,612	0,653	0,669	0,689	0,684	0,674	0,674
7	0,683	0,684	0,634	0,607	0,664	0,691	0,689	0,678	0,695	0,672
8	0,689	0,675	0,630	0,614	0,668	0,673	0,687	0,692	0,679	0,680
9	0,681	0,673	0,636	0,631	0,673	0,697	0,685	0,693	**0,708**	0,673
10	0,701	0,677	0,643	0,629	0,665	0,680	0,683	0,698	0,690	0,674
11	**0,703**	**0,694**	0,628	0,643	**0,692**	0,686	0,679	0,683	0,676	0,668
12	0,699	0,680	0,653	0,637	0,675	0,682	0,675	**0,701**	0,676	0,694
13	0,690	0,689	0,642	0,642	0,689	0,688	0,687	0,687	0,688	0,686
14	0,690	0,685	0,636	0,639	0,677	0,683	0,680	0,688	0,686	0,673
15	0,699	0,684	0,653	0,647	0,661	0,677	0,702	0,700	0,689	0,666
16	0,692	0,685	0,655	0,649	0,664	0,678	0,692	0,692	0,679	0,701
17	0,678	0,685	0,651	0,654	0,683	**0,692**	0,673	0,685	0,689	0,684
18	0,699	0,677	0,646	0,633	0,682	0,680	0,700	0,686	0,693	0,686
19	0,678	0,683	0,650	0,637	0,682	0,687	**0,717**	0,682	0,687	0,692
20	0,682	0,689	0,653	0,654	0,665	0,692	0,681	0,688	0,683	0,689
21	0,689	0,678	0,641	0,650	0,676	0,688	0,682	0,685	0,673	0,674
22	0,691	0,677	0,649	0,650	0,667	0,688	0,690	0,697	0,693	0,677
23	0,687	0,686	0,658	0,660	0,674	0,700	0,690	0,687	0,681	0,675
24	0,689	0,680	0,662	0,648	0,667	0,684	0,680	0,681	0,676	0,689
25	0,679	0,675	**0,670**	**0,680**	0,677	0,700	0,713	0,691	0,678	**0,691**

Tabelle 5.12.: Lernverlauf CNN, F-Score, gerundet

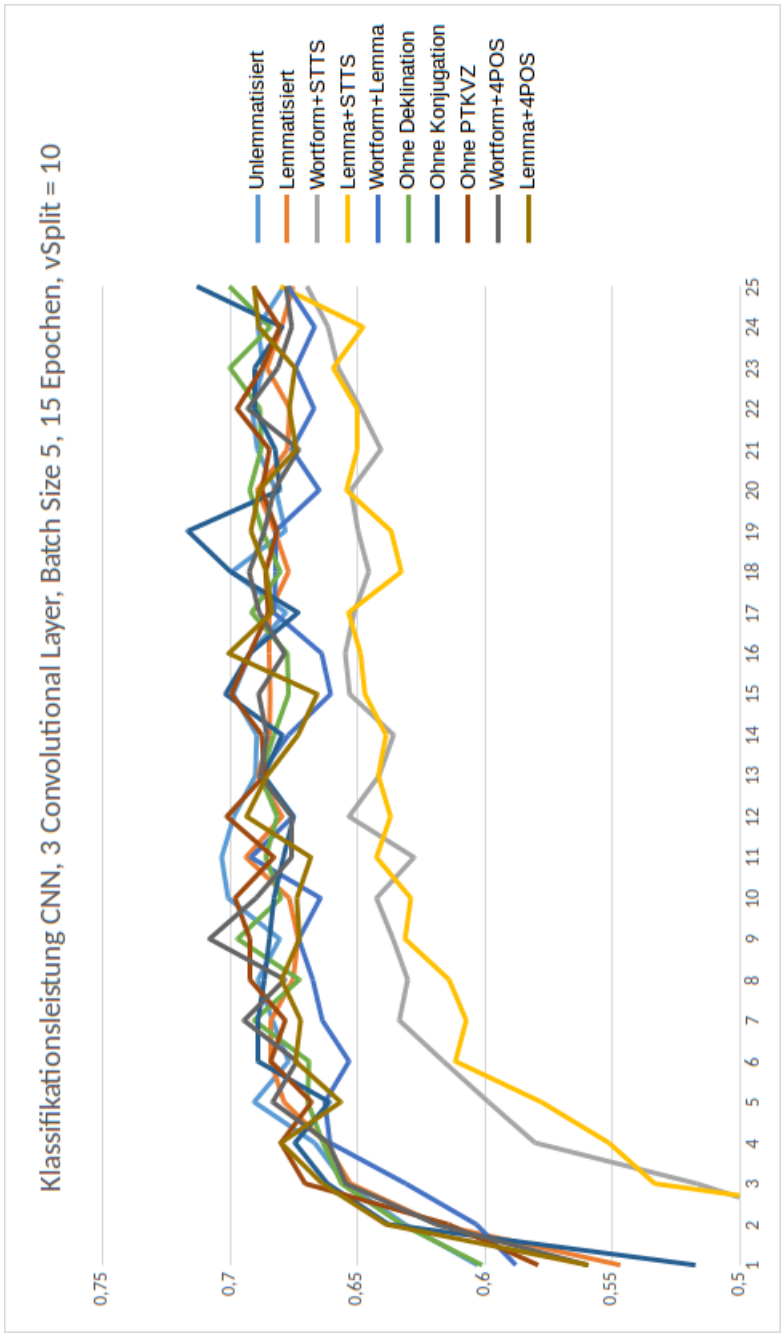


Abbildung 5.28.: Klassifikationsergebnis CNN, 3 Hidden Layer, Batch Size 5, Lerndauer 15 Epochen, 40x kreuzvalidiert, F-Score zu Trainingsmenge

Aus der Tabelle und der Visualisierung ersichtlich ist die auch nach 40facher Kreuzvalidierung verbleibende große Varianz der Modelle, die zu zahlreichen Überschneidungen untereinander und dem Effekt führt, dass häufiger als bei den konventionellen Modellen die besten F-Scores nicht notwendigerweise bei vollständigem Trainingskorpus erreicht werden. Augenscheinlich sind die niedrigen Einstiegswerte und flacheren Lernkurven der beiden STTS-erweiterten Modelle. Diese Modelle fügen wie im vorherigen Unterabschnitt beschrieben den Wortvektoren eine zusätzliche Dimension zur Notation des STTS-Tags als Ganzzahl hinzu; eine Information, die für den Klassifikator über weite Strecken des Lernvorgangs nur schwer zu modellieren scheint. Die erfolgreicheren übrigen Modelle hingegen demonstrieren mit einer steilen Lernkurve die Fähigkeit, bereits um den fünften Korpusgrößenpunkt, entsprechend 20% des Trainingskorpus und damit um bis zu 50% schneller als die konventionellen Modelle der Vierergruppe ihre Leistungsspitzenwerte anzusteuern.

Die aus den 40 Testläufen ermittelten Bestwerte der auf den verschiedenen Korpusversionen trainierten Modelle und die jeweils zugrundeliegenden Einzelwerte wurden einem Zweistichproben-T-Test mit dem Höchstwert der unlemmatisierten Originalkorpusversion unterzogen. Zur Untersuchung des Einflusses einer Korpusmodifikationsmaßnahme zu prüfen war die Nullhypothese H_0 übereinstimmender Mittelwerte des Modells *Unlemmatisiert* mit der jeweiligen modifizierten Version. Tabelle 5.13 zeigt, dass auf einem Signifikanzniveau von $p<0{,}05$ lediglich die Korpora *Wortform+STTS* und *Wortform+4POS* signifikant nach unten beziehungsweise oben vom besten Mittelwert der unlemmatisierten Originalversion abweichen.

Die erstgenannte Korpusversion teilt wie bereits ermittelt die schwer zu erlernende 301. Dimension für das STTS-Tagset mit ihrem lemmatisierten Pendant, kann aber, mutmaßlich aufgrund der nochmals erhöhten Symbolmenge, diesen Rückstand bis zum Ende des Trainingsvorgangs zumindest auf der vorliegenden Korpusgröße nicht wettmachen. Moderat, aber signifikant stärker als das Originalkorpus schneidet lediglich die Korpus-

Korpus	Mittelwert	Standardabweichung	p-Wert
UL	0,70325	0,053870	-
L	0,69375	0,051705	0,429289
WF+L	0,69175	0,052625	0,866002
WF+STTS	0,66975	0,050916	**0,005826**
L+STTS	0,68025	0,061053	0,411980
OD	0,70025	0,046286	0,451090
OK	0,71650	0,033132	0,078511
OP	0,70125	0,047127	0,102315
WF+4POS	0,70800	0,043715	**0,023656**
L+4POS	0,70050	0,049292	0,479264

Tabelle 5.13.: Klassifikationsergebnisse CNN, 40x kreuzvalidiert, mit p-Werten

version *Wortform+4POS* ab. Die in einem Unigrammvektor fehlende Möglichkeit, möglicherweise latent enthaltene Informationen über vorhandene Homographien über den Verwendungskontext aufzulösen, erscheint als denkbare Ursache dafür, dass allein diese Korpusmodifikation dem neuronalen Netz eine moderate Entlastung zu bieten scheint.

5.3.4. Untersuchungsklassifikatoren zu Embeddings

Das im vorherigen Unterkapitel zur Klassifikation eingesetzte neuronale Netz zeigte über den erstplatzierten F-Score im Vergleich zu sämtlichen konventionellen Klassifikatoren hinaus weitgehende Unempfindlichkeit gegenüber den untersuchten Flexionsphänomenen. Diese Unempfindlichkeit kann aus dem Ausbleiben von Leistungssteigerungen unter Einsatz der modifizierten Korpora geschlossen werden. Diese beobachtete Regelmäßigkeit schließt aus, dass die leicht erhöhte Klassifikationsleistung des neuronalen Netzes lediglich aus größerer formaler Ausdrucksstärke, semantischen Informationen im verwendeten Embedding oder einer Kombination von beidem resultierte. Dem Designprinzip vortrainierter Embeddings wie FastText entsprechend finden sich in den vortrainierten

Vektoren neben semantischen auch morphologische Informationen in Form latenter Variablen.

Ergänzend werden in diesem Ausschnitt ausgewählte Untersuchungen zu einzelnen Flexionsaspekten der offenen Wortklassen in Form binärer Klassifikationsaufgaben präsentiert. Zugrundeliegende Annahme ist hier, dass das neuronale Netz mit vergleichsweise geringem Architektur- und Trainingsaufwand imstande sein sollte, Flexionsmerkmale als für eine Themenklassifikation irrelevant einzustufen und entsprechend herunterzugewichten. Zu diesem Zweck wurden einem einfachen neuronalen Netz Vektoren jeweils zweier Wortformen präsentiert, mit dem Klassifikationszweck der Bestimmung einer morphologischen Beziehung dieser Wortformen. Da ein unigrammbasiertes Embedding wie FastText inhärent nicht in der Lage ist, Homographien kontextbasiert aufzulösen, diese im Trainingsvorgang jedoch in den Vektor einfließen sollten, wurde in einem separaten Experiment untersucht, ob aus dem Embeddingvektor einer Wortform zumindest auf das Vorhandensein von Homographien zu anderen Wortklassen geschlossen werden kann.

Das Format binärer Klassifikationsentscheidungen, etwa, ob ein Infinitiv eines Verbs in Relation zu einer konjugierten Verbform steht und es sich beispielsweise um dessen 1. Person Singular Indikativ Präsens Aktiv handelt, wurde gewählt, da es effizient zu berechnen ist und da die resultierende Aussage sich in Form einer einzigen Einheit speichern und in einer komplexeren Architektur weiterverwenden ließe. Alternativ denkbar wäre zum einen eine Reihe weiterer Klassifikationsformate: Statt einer binären Entscheidung, ob es sich bei einer Wortform um eine konjugierte Form oder einen Infinitiv einer Verbform handelt, wäre beispielsweise das Benennen einer konjugierten Form zu einem Infinitiv aus einer Menge verfügbarer Wortformen als Klassifikationsziel zu leisten. Eine weitere Alternative wäre das Benennen kategorialer Merkmale einer präsentierten Wortform, etwa Numerus, Tempus, Person oder eine Kombination mehrerer Kategorien. Die Abschnitte zu den einzelnen Klassifikationsexperimenten konkretisieren diese möglichen Alternativen zum jeweiligen Experiment.

Sämtlichen im Folgenden vorgestellten Untersuchungsklassifikatoren gemein ist der Verzicht auf jegliche Hidden Layer, das heißt, die Eingabe-Embeddings wurden unmittelbar mit der binären, sigmoidaktivierten Ausgabeeinheit verbunden. Die Einfügung eines oder mehrerer Hidden Layer erbrachte in keinem Fall eine Leistungssteigerung, ein angesichts des Designprinzips des Embeddings wenig überraschendes Ergebnis. Jedes Experiment wurde achtfach kreuzvalidiert; bei den präsentierten Ergebnissen handelt es sich um die Mittelwerte dieser jeweiligen Testserien.

5.3.4.1. Konjugation

Exemplarisch für die Kondensation von Konjugationsmerkmalen wurde das folgende Experiment durchgeführt: Dem neuronalen Klassifikator wurden in fixer Position die Vektoren eines Infinitiv und eines Verbs in der 1. Person Singular Indikativ (Aktiv) Präsens gezeigt, wobei keine Unterscheidung zwischen starken und schwachen Verben getroffen wurde. Das Klassifikationsziel bestand in der Entscheidung über die Zusammgenhörigkeit beider Wortformen, somit, ob es sich bei dem ersten Vektor um die Repräsentation des Infinitivs der im zweiten Vektor repräsentierten konjugierten Form handle. Während die Zusammengehörigkeit der Vektoren zu *aufen* und *kaufe* in diesem Sinne zu bejahen war, erfüllte eine Kombination etwa der Vektoren *verlassen* und *laufe* diese Anforderung nicht. Abbildung 5.29 zeigt die Lernkurve des Klassifikators als Funktion der Trainingsmenge.

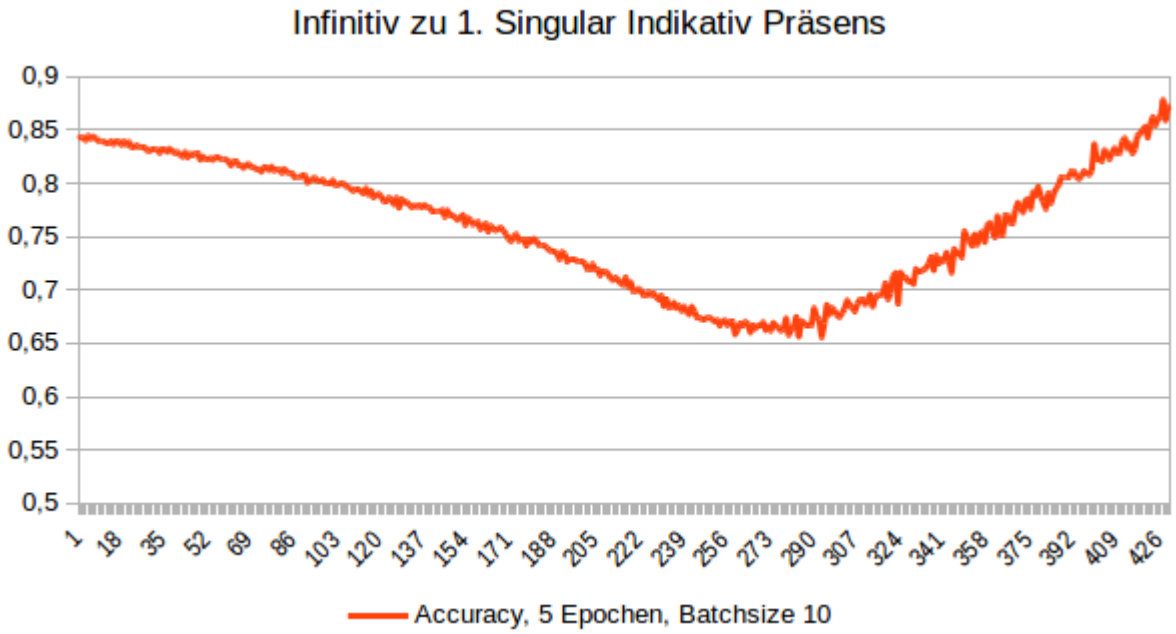


Abbildung 5.29.: Lernkurve Infinitiv und 1. Person Singular Indikativ Präsens Aktiv, Accuracy zu Trainingsbeispielen x 10

Nach einem Einstieg knapp unter 0,85 sinkt die Ergebniskurve zunehmend steil auf einen Sattelpunkt knapp oberhalb von 0,65 bei gut zwei Dritteln der Trainingsmenge, um dann näherungsweise linear auf Werte oberhalb von 0,85 zu steigen, wobei eine Konvergenzbewegung zum Ende der verfügbaren Trainingsbeispiele nicht abzusehen ist. Die Ursache für diesen temporären Abfall wurde nicht weiter untersucht. Eine denkbare Erklärung liegt jedoch im unterschiedlichen Valenzrahmen der zufallsgesampelten Verben: FastText-Vektoren beruhen in letzter Konsequenz auf räumlicher Distribution der vektorisierten Wortformen. Im Fall von Verben können unterschiedliche fakultative wie obligatorische Argumenterfordernisse die Kodierung der Wortklasse und nachfolgend verbalkategorialen Werte nach Dimensionen und Wertebereichen gruppieren. Die Beobachtung verschiedener Beispielverben aus unterschiedlichen Valenzgruppen im Training sorgt möglicherweise für Konfusion bei der notwendigen Gewichtung verbspezifischer

Informationen im Vektor durch das lernende Netz. Diese Probleme werden möglicherweise im Folgenden langfristig durch das Auftreten weiterer Vertreter der jeweils gleichen Valenzklasse kompensiert, so dass ein erlernbares gruppenbildendes Gewichtungsregime das Ergebnis über den Ausgangszustand hinaus in Richtung der Accuracy von $0,9$ entwickelt.

Zu diesem Experiment im Speziellen und dem Phänomen der Konjugation im Allgemeinen bieten sich zahlreiche ergänzende Experimente an: In einem ersten Schritt kann der Vergleich des Infinitivs mit jeder konjugierten Verbform durchgeführt beziehungsweise die Feststellung der Verwandtschaft beider Verbformen in binärer Klassifikation etabliert werden. Als Klassifikationsziel denkbar ist ferner jede mögliche Kombination der Verbformen untereinander. Diese Klassifikation wiederum muss nicht notwendigerweise innerhalb lediglich einer verbalen Kategorie erfolgen (beispielsweise „handelt es sich bei *gingt* um die 2. Person Plural Präteritum Indikativ Aktiv der 2. Person Singular Präsens Indikativ Aktiv *gehst*?“). Aus den Kategorien des Verbs und ihren möglichen Belegungen ergibt sich somit bereits eine Vielzahl möglicher binärer Klassifikationsexperimente. Diese Klassifikation kann sodann auch auf das Partizip II und möglicherweise auch auf das in dieser Arbeit als adjektivisch betrachtete Partizip I ausgedehnt werden. Weitere binäre Klassifikationsaufgaben schließen Konstellationen wie die Frage nach der Zugehörigkeit mehrerer Verbformen zueinander oder zum selben Infinitiv ein, etwa der 1. bis 3. Person Indikativ Aktiv in beliebigem Numerus oder Tempus („Gehören *gehe*, *gehst* und *geht* zueinander; handelt es sich um Formen des selben Infinitivs *gehen*?“). Schließlich kann die Art des Klassifikationsziels von binär auf multikategorial erweitert werden, etwa durch Auflistung aller Präsensformen, Benennung des korrekten Präteritums oder Plurals aus einer Menge von Vektoren.

5.3.4.2. Abtrennbare Verbpartikeln

Das im Folgenden vorgestellte Experiment prüfte die Zugehörigkeit einer abtrennbaren Verbpartikel zu einem mit dieser Partikel erweiterten Infinitiv, beispielsweise *ein*0 zu *einkaufen* oder *auf* zu *aufkündigen*. Die Darstellung der Ergebnisse in Abbildung 5.30 erfolgt exemplarisch mit den Lernkurven dreier unterschiedlich langer Trainingszeiträume von 5 bis 15 Epochen. Alle drei Kurven weisen bei ebenfalls acht Kreuzvalidierungen eine höhere Varianz auf als das vorhergehende Experiment. Tendenziell scheint sich jedoch auch dieser Klassifikator einer Unterscheidungsfähigkeit von 0,90 zu nähern.

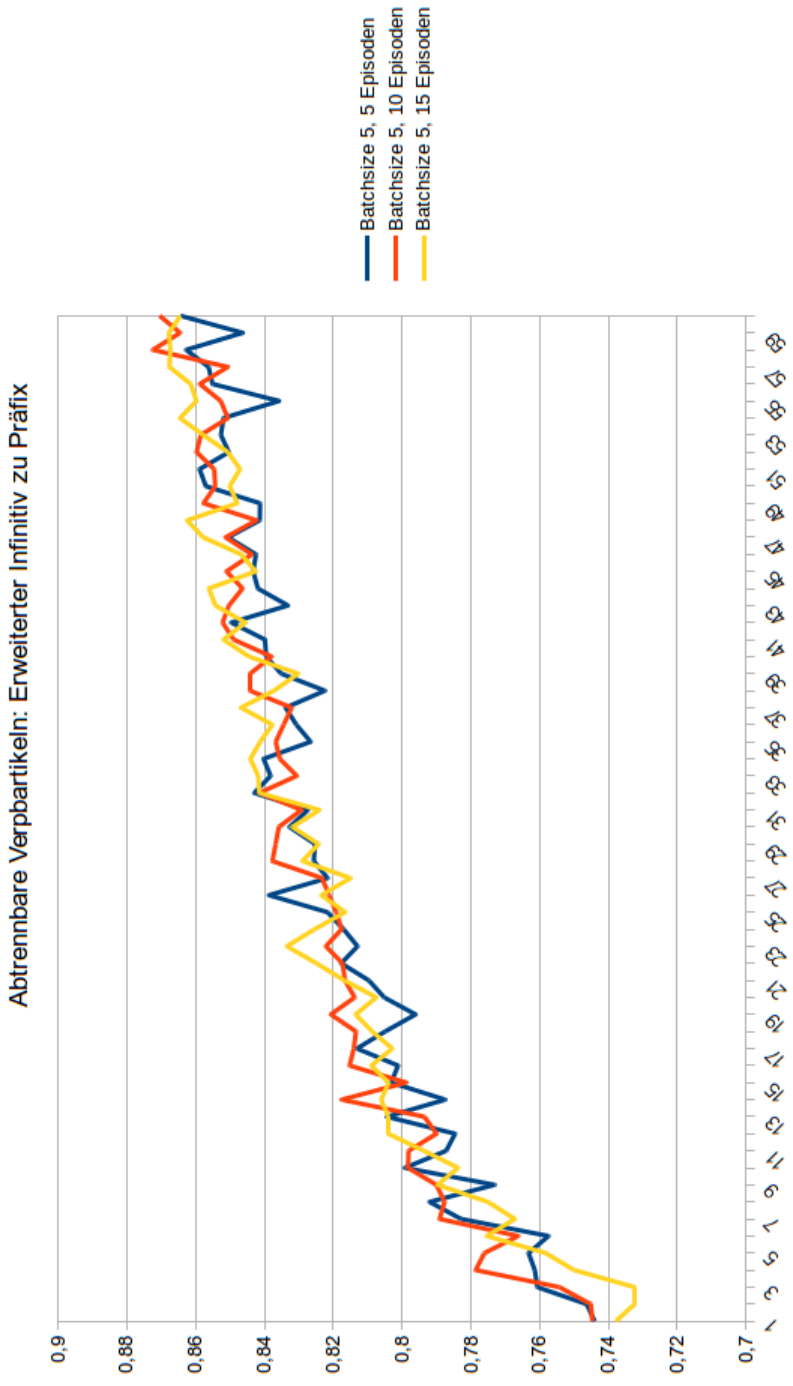


Abbildung 5.30.: Lernkurve erweiterter Infinitiv und Partikel, Accuracy zu Trainingsbeispielen x 10

Alternativ zur Prüfung der Fähigkeit einer Partikel, Teil eines Infinitivs wie in oben stehenden Beispielen zu sein, erscheinen zumindest zwei verwandte Experimente möglich: Die Prüfung der Möglichkeit einer Partikel, einen Infinitiv sinnvoll zu erweitern, etwa *an* den Infinitiv *rufen* (zu *anrufen*, eher nicht jedoch zu *anschlafen*, sowie die Frage, ob ein zugrundeliegender Infinitiv im Sinne einer Zeichenkette Grundlage eines solchermaßen erweiterten Infinitivs, etwa *steigen* für *aussteigen* (nicht jedoch für *ausmalen* oder *abkochen*) sein kann.

Letzteres Experiment könnte im Fall fakultativ abtrennbarer Partikeln mit semantischem Unterschied (etwa: *überspringen – springt über* vs. *überspringt*, *umfahren – fährt um* vs. *umfährt*) interessante Erkenntnisse über die internen Repräsentationen dieser homographieähnlichen Ambiguitäten in den Vektoren zutagefördern.

5.3.4.3. Homographie

Das folgende Experiment prüft stellvertretend für wortklassenübergreifende Homographie das Vorliegen einer solchen zwischen einem Substantiv und einem Adjektiv, unabhängig vom Vorliegen als Lemmata oder in deklinierter Form. Dem Klassifikator wird der Vektor einer homographen Wortform, etwa *wüste*, vorgelegt, bei der eine solche wortklassenspezifische Homographie zu bejahen ist, oder der einer nichthomographen Wortform, etwa *portfolio*. Abbildung 5.31 zeigt eine konvex scheinende Lernkurve ähnlich derer des ersten Experiments, ohne dass am Ende der möglichen Trainingsmengenvergrößerungen die Eingangsleistungsstärke bereits wieder erreicht wurde.

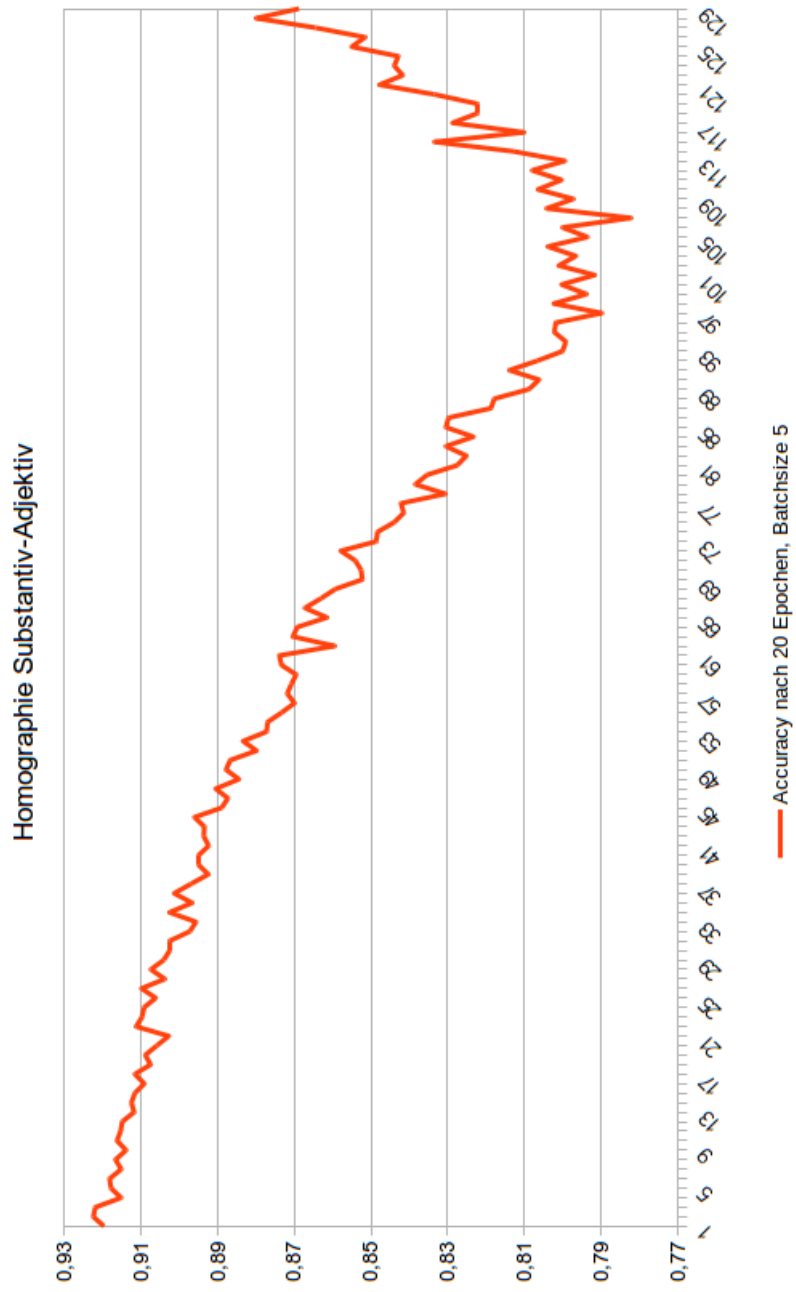


Abbildung 5.31.: Lernkurve Homographie Substantiv-Adjektiv, Accuracy zu Trainingsbeispielen x 10

Im Kontrast zum vorhergehenden Experiment zeigt sich hier bei gleichstarker Kreuzvalidierung eine wesentlich geringere Varianz. Analog zum ersten Experiment ist bei sich erholender Klassifikationsleistung allerdings eine deutlich gesteigerte Varianz im Vergleich zur ersten Hälfte des Trainingsverlaufs zu konstatieren.

Denkbare Ergänzungen zu diesem Experiment wären naheliegenderweise die Untersuchung der Sichtbarkeit von Homographie zwischen Substantiven und Verben sowie Adjektiven und Verben. Da die Vektoren der homographen Substantive/Verben ausweislich des durchgeführten Experiments Merkmale beider Wortklassen aufweisen, wäre des Weiteren ein Experiment vorstellbar, das eine Wortform zum Ziel einer Klassifikation in die vier Klassen *Substantiv*, *Verb*, *Homograph Substantiv-Verb* oder *Sonstiges* macht.

5.3.5. Zusammenfassung und Diskussion zur neuronalen Klassifikation

Die Experimente des Abschnitts 5.3.3 zeigen die Fähigkeit eines für heutige Verhältnisse architektonisch einfachen künstlichen neuronalen Netzes, unter Verwendung eines Unigram-Embeddings wie FastText die Ergebnisse der konventionellen Klassfiikatoren mit weniger als der Hälfte des bei diesen erforderlichen Trainingsaufwandes zu übertreffen. Dieser Erfolg beruht auf der Möglichkeit, das in den Embeddingvektoren über diese Wortformen enthaltene linguistische Wissen, das heißt das Wissen über ihre Semantik und Morphologie, wiederzuverwenden. Das im Zusammenhang mit dem Embeddingprinzip häufig zitierte „You shall know a word by the company it keeps“ (Firth, 1957) erinnert an das Designprinzip der Embeddings, numerische Vektoren aus der räumlichen Distribution von Wortformen aus einem Korpus zu extrahieren. Die in den Vektoren kondensierten Informationen entstammen der Analyse von Kollokationen in Korpora wie Wikipedia, die aus mehreren Milliarden Wortformen zu einer großen Bandbreite von Inhalten bestehen. Da die resultierenden Embeddingvektoren sowohl semantische

als auch morphologische Informationen enthalten, die durch das Convolutional Neural Network nach Bedarf unterschiedlich stark gewichtet werden können, erfolgt die Abgrenzung zwischen deren Anteilen am Klassifikationserfolg indirekt über die Verwendung der verschiedenen Korpusversionen: Das Ausbleiben jeglicher Verbesserung des Klassifikationserfolges bei Reduktion oder Entfernung der Flexionsphänomene durch Korpusmodifikationen stellt im Umkehrschluss den Nachweis dar, dass sämtliche benötigten Informationen hierüber bereits in den Vektoren enthalten sind und im Lernvorgang identifiziert werden können. Auf eine explizite Vorverarbeitung wie Lemmatisierung oder POS-Tagging des Korpus kann somit verzichtet werden. Dieser indirekte Nachweis mittels ausbleibender Verbesserung der Klassifikationsleistung durch Modifikationen wird expliziter und sichtbarer durch die Ergebnisse der Extraktionsexperimente in Abschnitt 5.3.4 gestützt: Ein künstliches neuronales Netz, dessen Eingabe lediglich aus zwei konkatenierten FastText-Vektoren besteht, ist selbst ohne verdeckte Schichten, also bei direkter Verbindung zur Ausgabeeinheit, in der Lage, mit hoher Treffgenauigkeit binäre Entscheidungen zu morphologischen Zusammenhängen zwischen Wortformen zu treffen. Starke Klassifikationsleistungen können in jeder der drei offenen Wortklassen und bereits nach wenigen hundert Beispielen erreicht werden. Selbst ohne komplexe Analyse interner Zustände des CNN-Nachrichtentextklassifikators erscheint denkbar, dass über die verdeckten Konvolutionsschichten zunächst eine implizite Lemmatisierung vorgenommen wird und sodann aus diesen Lemmata in späteren Schichten semantische Konzepte extrahiert werden. Die implizite Lemmatisierung ist möglicherweise nicht auf die im Training vorgelegten Lexeme beschränkt: Sind die Informationen der Flexionskategorien hinreichend kongruent in den Vektoren kodiert, erstreckt sich die erlernte Regelmäßigkeit, dass etwa das Tempus eines Verbs für den Klassifikationszweck nicht relevant ist, auf sämtliche Verbformen, indem die betreffenden Dimensionen generell entsprechend gewichtet werden. Somit würden nicht nur die Zusammenhänge zwischen bekannten Lexemen und ihren flektierten Formen implizit erlernt, sondern generell die zwischen allen Lexemen der jeweiligen Art. Vermeintlich unterstützende explizite Kodierung von

Part-of-Speech-Informationen in Form von diskretisierten STTS-POS-Tags scheinen den Lernvorgang des neuronalen Netzes eher zu belasten, während teilweise oder vollständige Lemmatisierung zumindest ohne Effekt bleibt. Die einzige statistisch verifizierbare Verbesserung scheint zu erfolgen, wenn das CNN durch die Ergänzung der Lemma-Vektoren um das kleine POS-Tag-Set explizit um das Einordnen der verbliebenen Homographen entlastet wird.

6. Diskussion und Ausblick

Die vorliegende Studie beschäftigte sich als erste ihrer Art gezielt mit dem Einfluss der Flexionsmorphologie der deutschen Sprache auf die konventionelle im Vergleich zur neuronalen automatischen Klassifikation von Texten nach Themen. Zu Beginn erfolgte eine systematische Darstellung der Flexionsvorgänge, die über die Bildung von Wortformen aus bedeutungstragenden Lexemen Einfluss auf Training und Betrieb von konventionellen wie neuronalen Klassifikatoren ausüben können. Dieser Darstellung und der Auflistung formloser Hypothesen zu Art und Ausmaß dieser Beeinflussung durch einzelne Phänomene folgte eine exemplarische empirische Überprüfung: Eine sechsstellige Anzahl sowohl konventioneller als auch neuronaler Klassifikationsexperimente wurde in einer großen Bandbreite von Parametrisierungen auf einem eigens für diese Untersuchung aus der Treebank TübaDZ geschaffenen annotierten Korpus von 3.643 deutschsprachigen Nachrichtentexten durchgeführt.

Die experimentellen Untersuchungen in dieser Arbeit ergaben eine Reihe deutlicher, teilweise von den Ausgangsyhypothesen abweichender Erkenntnisse. Zum einen erwies sich der Einfluss der Flexionsmorphologie insgesamt auf den Klassifikationsprozess als gering und nur teilweise signifikant sichtbar im Vergleich zu den Basisparametern Klassifikationsalgorithmus, Korpusgröße und Parameterkonfigurationen. Zum anderen entspricht die festgestellte Rangfolge der Einzelphänomene nicht den Annahmen, die sich aus der empirischen Korpusanalyse entwickeln ließen: Die ubiquitär auftretenden abtrennbaren

Verbpartikeln üben keinerlei sichtbaren Einfluss auf den Klassifikationserfolg aus. Die ebenso allgegenwärtige Homographie übt sichtbaren Einfluss erst in sehr großen Merkmalsräumen aus. Die formenreichen Verben üben geringeren Einfluss aus als die weit weniger morphologisch produktiven Substantive und Adjektive. Die Erstplatzierung der Deklination in der Rangfolge der Flexionsphänomene nach Einfluss scheint durch ihre quantitative empirische Stärke sowohl im Korpus als auch in den Merkmalsräumen und durch ihre gleichmäßiger verteilten Kategoriewerte begründet zu sein.

Zur Analyse der Vorgänge bei der Merkmalsraumbildung wurden drei umfangreiche Serien von Experimenten auf mit dem χ^2-Test gebildeten Merkmalsräumen verschiedener Größenordnungen durchgeführt. Das Experiment zur Merkmalsraumanalyse nach Wortklassen zeigte, dass die kleineren der aus den Nachrichtentexten extrahierten Merkmalsräume zu Beginn fast ausschließlich aus Substantiven bestehen. Erst mit zunehmender Größe werden auch die weiteren offenen Wortklassen bei einem steigenden Anteil von Homographen einbezogen, so dass auch andere Flexionsphänomene als Deklination korrespondierend in den Klassifikationsergebnissen eine Rolle spielen. Das Experiment zum Quotienten zwischen Lexemen und flektierten Formen zeigte, dass im Durchschnitt weit weniger verschiedene flektierte Formen als Merkmale akquiriert werden, als dies aus der Perspektive der Korpusebene zu erwarten gewesen wäre. Die Ursache hierfür ist mutmaßlich in der starken Konventionalisierung der Nachrichtentextsprache zu finden, wie sie besonders bei der Analyse der Verteilung von Person, Numerus und Tempus der Verben deutlich wird. Im dritten Experiment zur Merkmalsraumentwicklung schließlich werden die FastText-Embeddingvektoren aller auffindbaren Merkmale verschiedener Merkmalsräume miteinander verglichen, so dass ein Maß für die interne Dichte einer Kategorie in Form des durchschnittlichen Kosinusabstandes ihrer Merkmale verfügbar wird. Diese Metrik zeigt einen Verdichtungsprozess bei zunehmender Trainingskorpusgröße, in dem mutmaßliche morphologische Expansion mit semantischer Verdichtung interagiert. Da

FastText-Vektoren beide Dimensionen von Ähnlichkeit zwischen Wortformen abbilden können, ist der Anteil beider Phänomene nicht zu quantifizieren.

Die in Unterkapitel 4.6 aufgestellten Formalisierungen zeigen, dass eine Verdichtung von Dokumenten und Merkmalsräumen durch Lemmatisierung, also die Ausschaltung jeglicher Flexion, nicht pauschal erwartet werden kann: Lemmatisierung kann bei unterschiedlicher Verteilung flektierter Formen sowohl zum Ausschluss als auch zur Einbeziehung des selben Lexems aus dem Merkmalsraum führen. Dokumente, die im TF-IDF-Vektorformat mit der Standardmetrik Kosinusähnlichkeit verglichen werden, können einander durch Lemmatisierung ähnlicher oder unähnlicher werden. Diese Vorgänge beeinflussen unmittelbar vektorvergleichsbasierte Klassifikatoren wie Knn und Rocchio mutmaßlich stärker als lineare Klassifikatoren und künstliche neuronale Netze: Lineare Klassifikatoren können fehlerbehaftete Frequenzinformationen in Eingabevektoren möglicherweise durch die erlernten Gewichtungen kompensieren; künstliche neuronale Netze erhalten durch die Embeddingvektoren vielfältigere und tiefergehende Informationen zu den Wortformen des Textes, aus denen sie szenariospezifisch irrelevante Informationen, wie etwa Numerus und Tempus, durch Gewichtung herausfiltern können.

Die Ergebnisse der Klassifikationsexperimente scheinen die so gebildeten Annahmen zu stützen: Während die beiden vektorvergleichsbasierten Verfahren Knn und Rocchio zumindest im vorliegenden Szenario keine konkurrenzfähigen Klassifikationsleistungen erbringen können, handelt es sich bei den vier in gleicher Größenordnung klassifizierenden Verfahren (Logistische Regression, Naive Bayes Bi- und Multinomial und Support Vector Machine) um lineare Klassifikatoren. Die beste absolute Klassifikationsleistung und einen mehr als doppelt so schnellen Lernfortschritt erreicht konsequenterweise das embeddingbasierte Convolutional Neural Network.

Die Ermittlung des Einflusses der Flexionsphänomene erfolgte indirekt durch die Verwendung verschiedener Korpusmodifikationen, in denen Deklination, Konjugation inklu-

sive abtrennbarer Verbpartikeln und Homographie ganz oder teilweise entfernt wurden. Sämtliche Modifikationen erfolgten durch vollständige oder teilweise Lemmatisierung, Part-of-Speech-Tagging oder eine Kombination von Beidem. Die Verwendung zweier unterschiedlicher Tagsets (STTS vs. Kennzeichnung lediglich der offenen Wortklassen oder „Sonstiger“) zeigte, dass für ein optimales Klassifikationsergebnis ein Gleichgewicht zwischen der Symbolmenge, das heißt der Anzahl der erzeugten Merkmale, und dem Informationsgehalt pro Merkmal zu beachten ist. Im Szenario der Klassifikation von Nachrichtentexten nach Inhalten schneiden konsequenterweise Merkmale unter Verwendung des wortklassenbeschränkten Tagsets besser ab.

Die Erfolgsquote der Korpusmodifikationen folgt der Entwicklung der Anteile der Wortklassen in den Merkmalsräumen: In kleineren bis mittleren Merkmalsräumen werden Erstplatzierungen fast ausschließlich mit der deklinationsfreien Korpusmodifikation erreicht. Dieser Erfolg korrespondiert mit dem starken Übergewicht der Substantive und Adjektive im Merkmalsraum, die 80-100% der Merkmale konstituieren. Mit zunehmender Merkmalsraumgröße nehmen konjugierte Verbformen einen steigenden Anteil der Merkmale ein, so dass hier lemmatabasierte Modifikationen durch die zusätzliche Ausschaltung der Konjugation Gewinne erzielen und die deklinationsfreie Korpusmodifikation verdrängen können. In den größten Merkmalsräumen schließlich spielen Homographien eine derart sichtbare Rolle, dass die lemmatabasierte Modifikation mit kleinem POS-Tag-Set schließlich das beste Ergebnis erzielt, da sie sämtliche Flexion eliminiert und zusätzlich auch die in den Lemmata verbliebenen Homographien auflöst.

Ist der Einfluss der Flexion in den Experimenten zur konventionellen Klassifikation auch nach absoluten Werten überschaubar und nicht durchgehend statistisch gesichert, ist er dennoch sichtbar, folgt klaren, mit der vorgelagerten Merkmalsraumbildung korrespondierenden Trends und lässt sich nach nachvollziehbaren Mechanismen durch eindeutige Korpusmodifikationen verringern. Diese Beobachtungen können im embeddingbasierten Convolutional Neural Network hingegen nicht repliziert werden: Aus dem Ausbleiben

von Verbesserungen beim Klassifikationserfolg bei der Verwendung der flexionsreduzierten Korpusversionen kann geschlossen werden, dass das neuronale Modell keinerlei Empfindlichkeit gegenüber den primären Flexionsphänomenen aufweist. Verbesserungen werden lediglich bei der Verwendung der Korpusversion *Lemma+4POS* erreicht, was angesichts des Designs von FastText als Unigramm-Embedding nachvollziehbar erscheint: Jede Wortform wird durch genau einen Vektor abgebildet, der somit flektierte Formen separat formulieren kann, dabei aber Synkretismen und zufällige und flexionsinduzierte Homographien zusammenzieht. Das zweite signifikant abweichende Ergebnis, das schlechte Abschneiden der Korpusversion *Wortform+STTS*, lässt sich vermutlich dadurch erklären, dass redundante und somit überflüssige Informationen zur Wortklasse in schwer zu erlernender Form (301. Vektordimension mit dem Index eines STTS-Tags) eingeordnet werden müssen. Es handelt sich hierbei also mutmaßlich um eine technisch unbefriedigende Umsetzung des POS-Tag-Prinzips, die ohnehin keine klassifikationsrelevanten Zusatzinformationen kodiert.

Die sich an die Klassifikationsexperimente anschließenden Untersuchungsklassifikatoren (siehe Abschnitt 5.3.4) zeigen, dass Informationen über Flexionskategorien ebenso wie das Potenzial von Homographien in den Embeddingvektoren gespeichert und für ein simples Feed-Forward-Netz in Form einer binären Klassifikationsaufgabe leicht extrahierbar sind, etwa die Zusammenhänge zwischen konjugierten Verbformen und ihrem Lemma sowie abtrennbaren Partikeln und ihrem Stammverb. Die Fähigkeit eines solchen binären Klassifikators, das Vorhandensein von wortklassenübergreifenden Homographen zu erkennen, lässt es auch möglich erscheinen, ein solches Merkmal im Klassifikationsprozess konservativer zu gewichten.

Diese Arbeit traf eingangs eine Reihe von Annahmen zur linguistischen und technischen Eingrenzung ihres Untersuchungsgegenstandes, die im Folgenden zusammenfassend aufgezählt und möglichen Alternativen und Ergänzungen gegenübergestellt werden.

Auf linguistischer Ebene erfolgte zunächst eine Eingrenzung der untersuchten Merkmale auf wortformenbasierte Unigramme, das heißt einzelne, flektierte oder lemmatisierte Wörter. Prinzipiell sind als Merkmale aber auch Bi-, Tri- oder beliebige N-Gramme und deren jeweilige Erweiterungen etwa um POS-Tags möglich. N-Gramme können beispielsweise zur Auflösung von Homographen genutzt werden (*schmackhaften weinen* vs. *wir weinen*). Darüber hinaus eignen sie sich auch zum Auflösen von Synkretismen (*er weint* vs. *ihr weint*), sollten sich diese als klassifikationsrelevant erweisen. N-Gramme unterliegen jedoch dem sogenannten „Fluch der Dimensionalität“: Längere N-Gramme könnnen tendenziell informativer sein, treten allerdings seltener im Korpus auf. Die gut 146.000 Types der unlemmatisierten Originalversion des Nachrichtentextkorpus sind in der Lage, mehr als 21 Milliarden Bigramme zu bilden, von denen bei rund 1,8 Millionen Tokens selbst theoretisch maximal 0,01 Prozent im Korpus überhaupt auftreten können. Diese Tendenz verschärft sich exponentiell mit zunehmender Länge der N-Gramme und bleibt auch nach einer Lemmatisierung problematisch.Die Entscheidung für Wortformen-Unigramme und ihre lemmatisierten oder POS-Tag-ergänzten Pendants in den modifizierten Korpusversionen ist bereits Folge einer noch grundsätzlicheren Entscheidung, überhaupt Wörter als Merkmalsbasis zu nehmen. Bereits Cavnar et al. (1994) demonstrieren die alternative Verwendung von buchstabenbasierten N-Grammen für Klassifikationszwecke (in diesem Fall neben Inhaltskategorien auch für die Klassifikation nach der Sprache eines Dokuments). N-Gramme von Buchstaben können indirekt über die Darstellung von Affixen morphologieähnliche Modellierungskapazitäten aufweisen. Sie können auch Umlautungen teilweise kompensieren, da die N-Gramme, die den Umlaut nicht enthalten, für die Formen identisch sind. Des Weiteren wurde die pauschale Annahme getroffen, dass die Flexion der inhaltstragenden Lexeme keinerlei Korrelation mit dem Inhalt des Dokuments und den Eigenschaften der Kategorie aufweist. Hierzu erscheint denkbar, dass der Numerus von Verben oder die kongruenzbedingten Flexionsendungen von Adjektiven in Nominalgruppen durchaus vom Inhalt der Kategorie beeinflusst werden.

Sämtliche Korpusanalysen basierten auf dem in TübaDZ bereitgestellten Goldstandard zu POS-Tags und Flexionsinformationen. Auch die Modifikationen des Korpus für die Klassifikationsexperimente basierten auf der Verwendung dieser goldstandardgesicherten Informationen. In einem realen Klassifikationsszenario ist jedoch in aller Regel keine Goldstandardannotation zu diesen Eigenschaften vorhanden, so dass Part-of-Speech-Tags und Lemmata mittels automatischer Verarbeitung ermittelt werden müssen. Giesbrecht and Evert (2009) unterziehen eine Reihe damaliger State-of-the-Art-Tagger einer kritischen Replikationsstudie. Während sie im Wesentlichen die Größenordnung der von Brill (1992), Schmid (1999), Brants (2000) und Toutanova et al. (2003) gemeldeten Accuracy-Werte oberhalb von 97% für domäneninternes POS-Tagging bestätigen können, bemerken sie einen signifikanten Abfall beim Test auf einem domänenfremden Web-Crawl-Korpus auf rund 92%. Bei sogenannten Out-of-Vocabulary-Wörtern (OOV), also unbekanntem Vokabular, fallen die Werte auf bis zu 84%. Selbst ohne diese eklatanten Leistungsverluste bei Verlassen der idealisierten Validierungsumgebung gilt jedoch beim Blick auf die durch POS-Tagging und Lemmatisierung erreichten Werte in dieser Untersuchung: Die Verbesserung der Klassifikationsergebnisse durch vollständige (Logistische Regression und Support Vector Machine) oder teilweise Lemmatisierung (Naive Bayes) beträgt bescheidene 1,94, 3,48, 2,16 und 2,66 F-Score-Punkte – Angesichts dieser geringfügigen Verbesserungen im Vergleich zur gesicherten und mutmaßlich noch erheblicheren Fehlermarge der verbreiteten POS-Tagger ist nicht davon auszugehen, dass eine vollautomatische Annotation überhaupt messbare Verbesserungen gegenüber dem unlemmatisierten Originalkorpus erreichen kann. Die Ergebnisverbesserungen ergeben sich überdies durchweg nur in größeren Merkmalsräumen, bei denen davon auszugehen ist, dass sie anteilig mehr seltene Merkmale mit niedriger Korpusfrequenz enthalten, wodurch mit zusätzlichen Schwierigkeiten bei POS-Tagging und Lemmatisierung im Vergleich zum Korpusdurchschnitt zu rechnen ist.

Eine weitere linguistische Einschränkung betrifft offensichtlich die Auswahl und Gestaltung des Korpus für die experimentellen Untersuchungen. Unterkapitel 3.1 begründet diese Auswahl im Vergleich zu etablierten Standardkorpora. Dennoch erscheint die Frage nach einer Übertragbarkeit der gewonnenen Erkenntnisse in andere Domänen und Register weiterhin interessant. Nachrichtentexte der 1990er-Jahre, deren Vokabular sich durch geänderte Themen von heutigen Nachrichtentexten unterscheiden muss und weder im Wiktionary-Extrakt noch in FastText vollständig aufzufinden ist, schneiden möglicherweise in diesem Zusammenhang schwächer ab als zeitgenössische Pendants. Prinzipiell sollten in dieser Arbeit festgestellte Effekte auch in anderen Registern, etwa Social Media-Texten, anzutreffen sein: Zipfs Gesetz und seine beschriebenen Implikationen gelten für sämtliche Textgenres, die wiederum durch eigene sprachliche Konventionen ähnliche Ungleichverteilungen einzelner Flexionskategoriewerte erzeugen können. Wenig wahrscheinlich erscheint hingegen eine Übertragbarkeit der hier gewonnenen Erkenntnisse etwa auf die computerlinguistischen Disziplin der *Authorship Attribution*, also Zuweisung von Autorenschaft.

Schließlich ließe sich eine Untersuchung wie die vorliegende grundsätzlich für jede flektierende Sprache durchführen und gewinnbringend mit Bag-of-Words-Modellen für isolierende Sprachen wie etwa Hawaiianisch vergleichen, die möglicherweise stärker von kontextmodellierenden Mechanismen wie N-Grammen oder den hier nicht behandelten Rekurrenten Neuronalen Netzen profitieren.

Das in dieser Untersuchung ausschließlich verwendete Merkmalsauswahlkriterium χ^2 kann durch in Unterkapitel 2.2 erwähnte Alternativen wie Mutual Information ersetzt werden. Deren Verwendung hat möglicherweise zu untersuchende quantitative Verschiebungen, jedoch mutmaßlich keinen vollständigen Bruch mit den hier beobachteten Mechanismen zur Folge: Prinzipiell beruhen statistische Merkmalsauswahlverfahren per Definition auf der Korrelation zwischen Symbol und Klasse und unterscheiden sich möglicherweise hauptsächlich im Hinblick auf quantitative Randbedingungen wie Hapaxle-

gomena. Bei Verwendung eines frequenzabhängigen Kriterums wie *Document Frequency* (Schütze et al. (2008)) wären hingegen Verben und unter diesen die starken Verben möglicherweise stärker von negativer Selektion aufgrund ungleich verteilten Auftretens einzelner Kategoriewerte betroffen.

Für das Frequenz-Gewichtungsmaß TF-IDF existieren alternative Normalisierungsmöglichkeiten sowie alternative Vergleichsverfahren für Dokumentvektoren untereinander (Manhattan-, Mahalanobis- und euklidische Distanz) (Schütze et al. (2008)).

Das verwendete künstliche neuronale Netz ist architektonisch kompakt und konnte aufgrund von Rechenkraftbeschränkungen nicht in jeder Parameterdimension ausführlich getestet werden. Diese Einschränkungen laufen dem Zweck der Untersuchung nicht zuwider, da bereits die Existenz eines einzelnen die konventionellen Klassifikatoren übertreffenden Modells nachweisen kann, dass embeddingbasierte Architekturen jenen bauartbedingt hinsichtlich Leistung und Lerngeschwindigkeit überlegen sind. Die höhere Lerngeschwindigkeit ergibt sich daraus, dass große Teile des Trainings in Form der linguistischen Informationen in den Embeddingvektoren bereits vorweggenommen wurden. Somit muss nur die aufgabenspezifische Interpretation bestimmter Aspekte dieses linguistischen Vorwissens trainiert werden. Nichtsdestotrotz erscheint denkbar, dass sich die Leistung der vorliegenden Architektur durch ausgedehntere Parametersuchräume verbessern ließe.

Die Arbeit an konventionellen Klassifikationsverfahren und konsequenterweise vorgelagerten, expliziten Merkmalsauswahlverfahren ist in den letzten Jahren mit dem Erfolg der embeddinggetriebenen neuronalen Klassifikationsverfahren und zahlreicher weiterer Embeddingsprachmodellanwendungen vollständig zum Erliegen gekommen. Eine spätere Wiederaufnahme der Arbeiten an Klassifikationverfahren, die ohne derartige vortrainierte und in Vektorform kondensierte Informationen auszukommen hätten, erscheint ausgeschlossen. Die Tagungsbände etwa der drei großen Konferenzen *Annual Meeting*

of the Association for Computational Linguistics (ACL) (Gurevych and Miyao (2018a), Gurevych and Miyao (2018b), Korhonen et al. (2019), Jurafsky et al. (2020)), *Empirical Methods in Natural Language Processing (EMNLP)* (Riloff et al. (2018), Inui et al. (2019), Webber et al. (2020)) und *International Conference on Computational Linguistics (COLING)* (Bender et al. (2018), Scott et al. (2020)) sowie die laufenden ergänzenden Veröffentlichungen der ACL[1] geben unmittelbaren Aufschluss über die vollständige Dominanz künstlicher neuronaler Netze in sämtlichen angewandten Teilgebieten der Computerlinguistik.

In jüngster Vergangenheit wurde die erste Generation kontextfreier Embeddings (siehe Unterabschnit 2.4.4.2) von den auf dem Attention-Prinzip zur Kontextmodellierung (Vaswani et al. (2017)) basierenden Transformermodellen BERT und GPT/GPT2 (Radford et al. (2018)) abgelöst. Aktuelle Arbeiten beziehen sich überwiegend auf zwei Aspekte zu diesen Modellen: Der Implementierung immer weiterer sprach- und endanwendungsspezifischer Spezialmodelle (sogenannte *Downstream Tasks*) und der zunehmend detaillierteren Untersuchung ihrer inneren, „tiefen“ Repräsentationen linguistischen und semantischen Wissens. Bei dieser Generation neuronaler Sprachmodelle dürfte es sich um die erste computerlinguistische Technologie handeln, bei der in einem derartigen Ausmaß erst *nach* ihrer Veröffentlichung und ihrem breiten Einsatz ein Verständnis ihrer Funktionsmechanismen erarbeitet wird. Die Modelle sind schlicht zu komplex und umfangreich, um unmittelbar nachvollziehbar zu sein. Kovaleva et al. (2019) berichten in einem Survey-Paper bereits von etwa 150 die Wirkungsweise von BERT untersuchenden Arbeiten. Bezeichnenderweise sprechen die Autoren vom „Current state of knowledge“ über die Arbeitsweise von BERT. Einen Überblick über Methodik und Designziele sogenannter *Probes* und *Control Tasks* zur Analyse tiefer neuronaler Modelle geben Hewitt and Liang (2019). Bei derartigen Probing-Klassifikatoren handelt es sich um komplexere Extraktionen linguistischen Wissens analog zu den in Abschnitt 5.3.4 vorgestellten

[1] https://www.aclweb.org/anthology/events/cl-2018/ -2019, -2020

FastText-Experimenten. Derart umfangreiche Untersuchungen scheinen erst mit der Etablierung der genannten Transformermodelle aufgekommen zu sein und für ältere Embeddings nicht in diesem Umfang vorzuliegen. Poerner et al. (2018) bieten einen Überblick über Probing-Ansätze vor der Zeit der Transformermodelle.

Dror et al. (2019) konstatieren darüberhinaus allgemeine Schwierigkeiten bei der Vergleichbarkeit aktueller Deep Learning-Modelle durch enorme und instransparente Schwankungen zufallsinitialisierter Modelle, die bereits in dem simplen, in dieser Arbeit verwendeten CNN-Klassifikator erschienen.

Grundsätzlich ist zu erkennen, dass neuronale Modelle linguistisches Wissen lediglich anders abbilden als ihre Vorgängerverfahren und durch das Vortrainieren auf riesigen Korpora akkumulieren können, um für spezialisierte Anwendungen kalibriert zu werden. Die in dieser Arbeit behandelten Phänomene sind somit keineswegs obsolet, sondern lediglich effektiver gelöst. So zeigen etwa Jawahar et al. (2019), dass BERT allgemeine Morphologie in lokalisierbaren tieferen Layern speichert, und Chen et al. (2019) analysieren, wie das Thema der Merkmalsauswahl indirekt in Form der Vokabularauswahl und -größe unter Rechenkapazitätsgesichtspunkten in neuronalen Netzen weiterexistiert. Tenney et al. (2019) kommen gar zu dem Schluss, dass BERT die klassische NLP-Pipeline bzw. -hierarchie „wiederentdecke“: Ihre Probing-Experimente zeigen geradezu hierarchisch angeordnete Layer für POS, Syntax und Semantik.

Zukünftige Herausforderungen bestehen neben der überbordenden Komplexität der Modelle, die innerhalb von weniger als drei Jahren im Verlauf der Arbeit an der vorliegenden Untersuchung die konventionellen Klassifikatoren ebenso wie beinahe sämtliche anderen konventionellen computerlinguistischen Werkzeuge ersetzt haben, in deren enormem Ressourcenverbrauch: Schwartz et al. (2019) beziffern die Steigerung des Berechnungsaufwands für die Erstellung konkurrenzfähiger Deep Learning-Modelle zwischen 2012 und 2018 bereits auf den Faktor 300.000 und weisen darauf hin, dass für Hardware- und

Energiekosten siebenstellige Beträge für Training und Evaluation eines Modells anfallen können. Dieser Ressourcenaufwand wird einerseits Einfluss auf den Kreis der Mitwirkenden an zukünftigen Entwicklungen ausüben, andererseits möglicherweise Bemühungen zum besseren Verständnis der linguistischen Modellierungen in Deep Learning-Modellen und ihrer effizienteren Nutzung motivieren (hierzu etwa Chaudhary et al. (2020)). Die vorliegende Arbeit stellt einen Betrag für den Ausbau eines solchen Verständnisses dar, da sie zeigt, inwiefern flexionsmorphologische Aspekte einer Sprache, hier der deutschen, bei der Klassifikation von Texten in dieser Sprache zu beachten sind und wie Klassifikationsverfahren mit ihnen umgehen.

Literaturverzeichnis

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16). pp. 265–283.

Bender, E. M., Derczynski, L., Isabelle, P. (Eds.), Aug. 2018. Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., 2003. A neural probabilistic language model. The journal of machine learning research 3, 1137–1155.

Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G., 2002. The tiger treebank. In: Proceedings of the workshop on treebanks and linguistic theories. Vol. 168.

Brants, T., Apr. 2000. TnT – a statistical part-of-speech tagger. In: Sixth Applied Natural Language Processing Conference. Association for Computational Linguistics, Seattle, Washington, USA, pp. 224–231.

Brill, E., Mar. 1992. A simple rule-based part of speech tagger. In: Third Conference on Applied Natural Language Processing. Association for Computational Linguistics, Trento, Italy, pp. 152–155.

Buchholz, S., Marsi, E., 2006. Conll-x shared task on multilingual dependency parsing. In: Proceedings of the tenth conference on computational natural language learning (CoNLL-X). pp. 149–164.

Cavnar, W. B., Trenkle, J. M., et al., 1994. N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. Vol. 161175. Citeseer.

Chaudhary, Y., Gupta, P., Saxena, K., Kulkarni, V., Runkler, T., Schütze, H., Nov. 2020. TopicBERT for energy efficient document classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp. 1682–1690.

Chen, W., Su, Y., Shen, Y., Chen, Z., Yan, X., Wang, W. Y., Jun. 2019. How large a vocabulary does text classification need? a variational approach to vocabulary selection. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3487–3497.

Chollet, F., et al., 2018. Keras: The python deep learning library. Astrophysics Source Code Library , ascl–1806.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20 (1), 37–46.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20 (3), 273–297.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory 13 (1), 21–27.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Dror, R., Shlomov, S., Reichart, R., Jul. 2019. Deep dominance - how to properly compare deep neural models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 2773–2785.

Eisenberg, P., 2020. Grundriss der deutschen Grammatik: Das Wort. J.B. Metzler, Stuttgart, Ch. Flexion, pp. 159–218.

Fabricz, K., 1986. Particle homonymy and machine translation. In: Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics.

Forman, G., et al., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar), 1289–1305.

Friedman, J., Hastie, T., Tibshirani, R., et al., 2001. The elements of statistical learning. Vol. 1. Springer series in statistics New York.

Giesbrecht, E., Evert, S., 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In: Proceedings of the fifth Web as Corpus workshop. pp. 27–35.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. Vol. 1. MIT press Cambridge.

Gurevych, I., Miyao, Y. (Eds.), Jul. 2018a. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia.

Gurevych, I., Miyao, Y. (Eds.), Jul. 2018b. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Melbourne, Australia.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., Seung, H. S., 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405 (6789), 947–951.

Hebb, D. O., 1949. The organization of behavior; a neuropsycholocigal theory. A Wiley Book in Clinical Psychology 62, 78.

Hewitt, J., Liang, P., Nov. 2019. Designing and interpreting probes with control tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 2733–2743.

Hull, D. A., Pedersen, J. O., Schütze, H., 1996. Method combination for document filtering. In: Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 279–287.

Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Nov. 2019. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y., 2009. What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th international conference on computer vision. IEEE, pp. 2146–2153.

Jawahar, G., Sagot, B., Seddah, D., Jul. 2019. What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 3651–3657.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Springer, pp. 137–142.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T., 2016a. Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 .

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016b. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 .

Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Jul. 2020. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online.

Kirov, C., Sylak-Glassman, J., Que, R., Yarowsky, D., 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3121–3126.

Klavans, J. L., Kan, M.-Y., Aug. 1998. Role of verbs in document analysis. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 680–686.

Korhonen, A., Traum, D., Màrquez, L. (Eds.), Jul. 2019. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy.

Kovaleva, O., Romanov, A., Rogers, A., Rumshisky, A., Nov. 2019. Revealing the dark secrets of BERT. In: Proceedings of the 2019 Conference on Empirical Methods in

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 4365–4374.

Krovetz, R., Jul. 1997. Homonymy and polysemy in information retrieval. In: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Madrid, Spain, pp. 72–79.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1 (4), 541–551.

Leopold, E., Kindermann, J., 2002. Text categorization with support vector machines. how to represent texts in input space? Machine Learning 46 (1), 423–444.

Lezius, W., Rapp, R., Wettler, M., 1998. A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In: COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics.

Lyons, J., 1990. Die Sprache. Beck.

Maas, A. L., Hannun, A. Y., Ng, A. Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. Vol. 30. Citeseer, p. 3.

Maron, M. E., Kuhns, J. L., 1960. On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM) 7 (3), 216–244.

McCulloch, W. S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5 (4), 115–133.

Meyer, C. M., Gurevych, I., 2010. Worth its weight in gold or yet another resourceâa comparative study of wiktionary, openthesaurus and germanet. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 38–49.

Meyer, C. M., Gurevych, I., 2012a. Ontowiktionary: Constructing an ontology from the collaborative online dictionary wiktionary. In: Semi-Automatic Ontology Development: Processes and Resources. IGI Global, pp. 131–161.

Meyer, C. M., Gurevych, I., 2012b. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. na.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 .

Oliphant, T. E., 2006. A guide to NumPy. Vol. 1. Trelgol Publishing USA.

Pachunke, T., Mertineit, O., Wothke, K., Schmidt, R., 1992. Broad coverage automatic morphological segmentation of German words. In: COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 .

Poerner, N., Schütze, H., Roth, B., Jul. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 340–350.

Powers, D. M. W., 1998. Applications and explanations of Zipf's law. In: New Methods in Language Processing and Computational Natural Language Learning.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training .

Ramachandran, P., Zoph, B., Le, Q. V., 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941 .

Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (Eds.), Oct.-Nov. 2018. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium.

Rocchio, J., 1971. Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing , 313–323.

Rogati, M., Yang, Y., 2002. High-performing feature selection for text classification. In: Proceedings of the eleventh international conference on Information and knowledge management. pp. 659–661.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65 (6), 386.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning representations by back-propagating errors. nature 323 (6088), 533–536.

Russell-Rose, T., Stevenson, M., Whitehead, M., 2002. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. .

Schiller, A., Teufel, S., Thielen, C., 1995. Guidelines f ur das tagging deutscher textcorpora mit stts. Universität Stuttgart, Universität Tübingen, Germany .

Schlippe, T., Ochs, S., Schultz, T., 2010. Wiktionary as a source for automatic pronunciation extraction. In: Eleventh Annual Conference of the International Speech Communication Association.

Schmid, H., 1999. Improvements in part-of-speech tagging with an application to german. In: Natural language processing using very large corpora. Springer, pp. 13–25.

Schütze, H., Hull, D. A., Pedersen, J. O., 1995. A comparison of classifiers and document representations for the routing problem. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 229–237.

Schütze, H., Manning, C. D., Raghavan, P., 2008. Introduction to information retrieval. Vol. 39. Cambridge University Press Cambridge.

Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O., 2019. Green ai. arXiv preprint arXiv:1907.10597 .

Schwenk, H., Li, X., 2018. A corpus for multilingual document classification in eight languages. arXiv preprint arXiv:1805.09821 .

Scott, D., Bel, N., Zong, C. (Eds.), Dec. 2020. Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online).

Sennrich, R., Kunz, B., 2014. Zmorge: A german morphological lexicon extracted from wiktionary .

Skut, W., Krenn, B., Brants, T., Uszkoreit, H., 1997. An annotation scheme for free word order languages. arXiv preprint cmp-lg/9702004 .

Telljohann, H., Hinrichs, E., Kübler, S., Kübler, R., 2004. The tüba-d/z treebank: Annotating german with a context-free backbone. In: In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004. Citeseer.

ten Hacken, P., Bopp, S., 1998. Separable verbs in a reusable morphological dictionary for German. In: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.

Tenney, I., Das, D., Pavlick, E., Jul. 2019. BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 4593–4601.

Thielen, C., 1999. Guidelines f ur das tagging deutscher textcorpora mit stts (kleines und gro es tagset) .

Toutanova, K., Klein, D., Manning, C. D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 252–259.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762 .

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods 17 (3), 261–272.

Weale, T., Brew, C., Fosler-Lussier, E., 2009. Using the wiktionary graph structure for synonym detection. In: Proceedings of the 2009 Workshop on The Peopleâs Web Meets NLP: Collaboratively Constructed Semantic Resources (Peopleâs Web). pp. 28–31.

Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Nov. 2020. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online.

Weber, H. J., 1973. The automatically built up homograph dictionary a component of a dynamic lexical system. In: COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics.

Wöllstein, A., 2016. Duden, Die Grammatik: unentbehrlich für richtiges Deutsch. Dudenverlag.

Yang, Y., Pedersen, J. O., 1997. A comparative study on feature selection in text categorization. In: Icml. Vol. 97. Nashville, TN, USA, p. 35.

Zesch, T., Müller, C., Gurevych, I., 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In: LREC. Vol. 8. pp. 1646–1652.

Zifonun, G., Hoffmann, L., Strecker, B., Ballweg, J., 1997. Grammatik der deutschen Sprache. Vol. 1. Walter de Gruyter.

Zipf, G. K., 1949. Human behavior and the principle of least effort: an introd. to human ecology .

A. Danksagung

An erster Stelle möchte ich mich bei meinem Forschungsgruppenleiter Prof. Dr. Ulrich Schade für die langjährige gute Betreuung bedanken, von der Vermittlung vieler Grundlagen der Computerlinguistik ganz am Anfang meines Werdegangs in unserem Fachgebiet bis zu den letzten Tagen der Anfertigung dieser Arbeit.

Mein großer Dank gilt Prof. Dr. Claudia Wich-Reif für die Betreuung und Begutachtung dieser Arbeit auf Seiten der Universität. Sie hat es mir mit Ratschlägen und Einschätzungen stets zum richtigen Zeitpunkt ermöglicht, dieses computerlinguistische Projekt um eine germanistische Perspektive erheblich zu bereichern.

Ich danke auch den weiteren Mitgliedern der Prüfungskommission, Prof. Dr. Kristian Berg als Vorsitzendem und Prof. Dr. Thomas Klein, für ihre Bereitschaft zur Abnahme meiner Disputation und ihre engagierten Beiträge.

Meinen Kolleginnen und Kollegen in der Forschungsgruppe Informationsanalyse am Fraunhofer FKIE, besonders Albert und Magdalena, möchte ich für ungezählte produktive wie gut gelaunte Diskussionen und Überlegungen im Rahmen unserer gesamten gemeinsamen Arbeit danken.

Mein größter Dank gebührt meiner Familie und besonders meinem Bruder Robert, und meinen Freunden: Dafür, dass sie meine Promotion als Station auf einem Weg, der sehr viel früher begonnen hat und noch nicht zuende ist, mit mir erlebt haben.