Originalveröffentlichung in: Journal for Language Technology and Computational Linguistics
Jg. 36 (2023) Nr. 1, S. 113-150.
DOI: *https://doi.org/10.21248/jlcl.36.2023.240*

Sarah Broll / Roman Schneider

Empirische Verortung konzeptioneller Nähe/Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora

Abstract

Linguistische Studien arbeiten häufig mit einer Differenzierung zwischen gesprochener und geschriebener Sprache bzw. zwischen Kommunikation der Nähe und Distanz. Die Annahme eines Kontinuums zwischen diesen Polen bietet sich für eine Verortung unterschiedlichster Äußerungsformen an, inklusive unkonventioneller Textsorten wie etwa Popsongs. Wir konzipieren, implementieren und evaluieren ein automatisiertes Verfahren, das mithilfe unkorrelierter Entscheidungsbäume entsprechende Vorhersagen auf Textebene durchführt. Für die Identifizierung der Pole definieren wir einen Merkmalskatalog aus Sprachphänomenen, die als Markierer für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit diskutiert werden, und wenden diesen auf prototypische Nähe-/Mündlichkeitstexte sowie prototypische Distanz-/Schrifttexte an. Basierend auf der sehr guten Klassifikationsgüte verorten wir anschließend eine Reihe weiterer Textsorten mithilfe der trainierten Klassifikatoren. Dabei erscheinen Popsongs als „mittige Textsorte", die linguistisch motivierte Merkmale unterschiedlicher Kontinuumsstufen vereint. Weiterhin weisen wir nach, dass unsere Modelle mündlich kommunizierte, aber vorab oder nachträglich verschriftlichte Äußerungen wie Reden oder Interviews vollkommen anders verorten als prototypische Gesprächsdaten und decken Klassifikationsunterschiede für Social-Media-Varianten auf. Ziel ist dabei nicht eine systematisch-verbindliche Einordung im Kontinuum, sondern eine empirische Annäherung an die Frage, welche maschinell vergleichsweise einfach bestimmbaren Merkmale („shallow features") nachweisbar Einfluss auf die Verortung haben.

Keywords: Mündlichkeit, Schriftlichkeit, Nähe, Distanz, Textsorten, Empirik, Features, Machine Learning

1 Motivation

Natürlicher Sprache begegnen wir üblicherweise in mündlich gesprochener (phonischer) oder geschriebener (graphischer) Form. Letztere nutzt analoge respektive digitale Kommunikationsmedien, erlaubt eine Trennung von Produktions- und Rezeptionsphase und mithin mannigfaltige Formen der Überarbeitung (Fehlerkorrektur, Stilprüfung etc.). Gesprochener Diskurs dagegen findet überwiegend spontan und ungeplant statt. Damit liegt nahe, „dass das gesprochene und das geschriebene Deutsch unterschiedliche Präferenzen bei der Wahl der Mittel aus dem Systeminventar des Deutschen haben"

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-122057

(Eichinger, 2017, S. 292). Äußerungen beider Modalitäten unterscheiden sich hinsichtlich Repertoire und Verwendung von syntaktischen Konstruktionen oder Vokabular, stilistischen Merkmalen – etwa Neubildungen oder Anakoluthen – und einigem mehr. Typisch lautliche Ausdrucksmöglichkeiten und Strategien auf der einen Seite kontrastieren mit einer (zumindest oft angenommenen) größeren Nähe zu (ebenfalls oft auch nur vermeintlichen) Sprachstandards und Konventionen.

Diese mediale Charakterisierung erscheint auf den ersten Blick als unkompliziert, erweist sich aber spätestens dann als unterkomplex, sobald alternative Kommunikationswege ins Spiel kommen: Songtexte beispielsweise lassen sich als medial mündlich kategorisieren, wenn sie live vorgetragen bzw. rezipiert werden, und als medial schriftlich, wenn wir uns auf ihre Darstellung in Liederbüchern o.Ä. beziehen. Gleiches gilt für viele andere kommunikative Formen wie Vorträge, Reden, Interviews usw., die geplant oder auch spontan stattfinden und vorab bzw. anschließend verschriftlicht werden.

Eine Überwindung des medialen Binarismus erlaubt der von Koch und Oesterreicher (1985, 2007) eingeführte und mittlerweile in der Linguistik weithin etablierte Ansatz eines multidimensionalen Kontinuums zwischen den beiden Polen 'Sprache der Nähe' und 'Sprache der Distanz'. Bei der Einordnung konkreter Äußerungen stehen hier nicht die Eigenschaften des Vermittlungsmediums im Vordergrund, sondern Kommunikationsbedingungen wie raumzeitliche Trennung, Reflektiertheit, Vertrautheit der Kommunikationspartner oder Affektivität. Aus diesen ergeben sich dann ggf. Präferenzen für die mediale Realisierung. Koch/Oesterreicher liefern zwar keine umfassende Systematik, welche sprachlichen Merkmale genau mit welchem Wirkungsgrad für eine Verortung im Nähe-Distanz-Kontinuum ausschlaggebend sind, gehen aber auf einige relevante universale und einzelsprachliche Eigenschaften ein, etwa morphosyntaktische Phänomene, lexikalische Vielfalt, Gliederungssignale, Verwendung von Partikeln usw. Eine klare und eindeutige Einordnung bleibt in vielen Fällen schwierig (vgl. z.B. Biber und Conrad (2019)).

Mit der vorliegenden Studie möchten wir keinen Beitrag leisten zu Debatten über Angemessenheit, Unzulänglichkeiten oder Modifikationen des Nähe-Distanz- bzw. Mündlich-Schriftlich-Ansatzes[1] (vgl. hierzu z. B. Feilke und Hennig (2016)), sondern die gewinnbringende Umsetzbarkeit seiner empirischen Algorithmisierung für die automatische Textklassifikation großer Datenmengen demonstrieren, basierend auf der Annahme eines konzeptionellen Kontinuums zwischen Nähesprache (mit einer Präferenz zur Mündlichkeit/Oralität) und Distanzsprache (mit einer Präferenz zur Schriftlichkeit/Literalität).

Digitale Infrastrukturen für Sprachressourcen spiegeln die medialen Pole wider: Die *Datenbank Gesprochenes Deutsch (DGD)* (Schmidt, 2017) als größte Sammlung gesprochener deutscher Sprache enthält ausschließlich Audioinhalte (und deren Transkripte), die mehr oder weniger typische mündliche Kommunikationssituationen repräsentieren. Im Gegensatz dazu umfasst das *Deutsche Referenzkorpus (DeReKo)* (Kupietz, Lüngen, Kamocki & Witt, 2018) medial schriftliche Texte, die dem Anschein nach in verschiede-

[1] In eine ähnliche Richtung zielen Unterscheidungen zwischen informeller und formeller Sprache oder zwischen Alltags-/Gebrauchs- und Bildungssprache; auf mögliche terminologische Überschneidungen und Unschärfen soll hier ebenfalls nicht eingegangen werden.

nen Bereichen der Nähe-Distanz-Kontinuumsskala angesiedelt werden können: Fach- und Publikumspresse, Belletristik, Redetranskripte, Soziale Medien etc. Die genaue Verortung ist keinesfalls trivial und konventionelle Metadaten helfen dabei nicht immer weiter: Zeitungen und Zeitschriften zum Beispiel umfassen zwar zuvorderst Beiträge, die vermutlich intuitiv als Distanzsprache (konzeptionell schriftlich) eingeordnet werden würden; sie enthalten aber ebenfalls Interviews, Gespräche und Diskussionen, die – ungeachtet der üblichen redaktionellen Nachbearbeitung und Fehlerkorrektur – tendenziell vermehrt Eigenschaften konzeptioneller Mündlichkeit aufweisen.

Offen bleibt angesichts der hohen Variationsbreite und Dynamiken bereits innerhalb etablierter Ressourcen die Klassifizierung korpuslinguistischer Spezialsammlungen wie der oben angesprochenen Songtexte: Tendieren diese zur Oralität oder zur Literalität oder sind sie irgendwie hybrid – und falls ja, aufgrund welcher Merkmale und mit welcher Gewichtung? Zuverlässig lassen sich solche Aussagen bestenfalls auf Einzeltextebene treffen, nicht pro Medientextsorte.[2] Für umfangreiche Datensamples existiert nichtsdestotrotz das Desiderat einer soliden automatisierten Verortung anhand empirisch ermittelbarer Kriterien.

Vor diesem Hintergrund verfolgt die vorliegende Studie drei Ziele: (a) Evaluation methodisch fundierter, datengesteuerter Klassifikationen konzeptioneller Mündlichkeit bzw. Schriftlichkeit auf einer belastbaren Textbasis (b) Operationalisierung der quantitativen Verortung auch nicht-eindeutiger Daten im Nähe-Distanz-Kontinuum (c) statistisch valide Identifikation wirkungsmächtiger sprachlicher Einflussfaktoren und Merkmale.

2 Related Work

Eine Reihe moderner Grammatik- bzw. Variationsbeschreibungen skizzieren Merkmale des gesprochenen Deutsch bzw. gehen auf umgangssprachliche Besonderheiten ein, exemplarisch seien Zifonun, Hoffmann und Strecker (1997) und Barbour und Stevenson (1998) genannt. Ágel und Hennig (2006a, 2006b) beziehen sich auf dem Weg zu einer Theorie des Nähe- und Distanzsprechens auf verschiedene von Koch/Oesterreicher angeführte Merkmale und konzipieren darüber hinausgehend eine differenzierte Systematik, die Nähemerkmale aus universalen Parametern der Nähekommunikation ableitet.

Wenige empirische Studien haben diese umfassenden Vorarbeiten bislang praktisch aufgegriffen und korpuslinguistisch evaluiert. Für das Deutsche berechnen Ortmann und Dipper (2019) eine Reihe einfacher sprachlicher Merkmale und nutzen diese zur Ermittlung des Grads an Oralität in Zeitungsartikeln (extrahiert aus Tüba-D/Z- und Tiger-Baumbanken), in Reden und Vorträgen (aus dem Gutenberg-DE-Korpus) sowie in weiteren Samples monologisch und dialogisch angelegter Diskurse (Predigten, Filmuntertitel, Gesprächs- und Chatprotokolle) aus öffentlichen verfügbaren Quellen bzw. Spezialkorpora. Das STTS-getaggte Korpus umfasst insgesamt ca. 2,5 Millionen Token; die 17 für die automatische Ermittlung herangezogenen Merkmale sind typisiert in die Kategorien Komplexität, Bezugnahme/Deixis, Syntax und Satztyp. Ortmann

[2] Und selbst dies gestaltet sich angesichts potenziell verschiedenartiger Textpassagen nicht durchgehend unproblematisch; vgl. (Dürscheid, 2016, S. 55).

und Dipper (2020) wenden den statistischen Ansatz mit Erfolg auch auf historische Texte an. Die kombinierten Erkenntnisse fließen in das Design eines übergreifenden, linguistisch motivierten Oralitäts-Maßes ein. [3]

Flankierend zu diesen Genre- und Textsorten-übergreifenden Arbeiten finden auf einzelne Medientypen fokussierte Auswertungen statt. Werner (2021) kontrastiert in einer multidimensionalen Registeranalyse (englischsprachige) Songtexte mit anderen Textsorten und weist trotz üblicherweise geplanter Textproduktion verschiedene mündlich/konversationelle Charakteristika nach. Androutsopoulos (2003) und Schlobinski (2005) beschreiben linguistische Features in Online-Sprache. Speziell für den Bereich der computervermittelten Kommunikation (Computer-Mediated Communication, CMC) untersucht Rehm (2002) die Verteilung eines kleinen Sets mutmaßlicher Merkmale konzeptioneller Mündlichkeit auf universitären Webseiten. Storrer (2000) thematisiert internetbasierte Kommunikationsformen wie E-Mail, Online-Foren und -Chats unter Bezugnahme auf entsprechende quantitative Erhebungen zur Einordnung in das Nähe-Distanz-Kontinuum. Für die Auswertung eines Chat-Korpus unterscheidet Kilian (2001) 14 linguistische Merkmale konzeptioneller Mündlichkeit, Cotgrove (2017) erweitert dieses Featureset im Rahmen einer Untersuchung von 600 deutschsprachigen Online-Kommentaren zu Youtube-Musikvideos.

3 Datengrundlage

3.1 Stratifikation des Untersuchungskorpus

Das Design unseres Untersuchungskorpus soll die beiden angenommenen Pole „konzeptionelle Mündlichkeit“ bzw. „konzeptionelle Schriftlichkeit“ unter Heranziehung einer aussagekräftigen Belegmenge authentischer Sprachsamples abdecken. Zudem dient es der experimentellen Verortung weiterer Sprachdaten aus verschiedenartigen Kommunikationskontexten. Zu diesen Zwecken fächern sich die Primärdaten auf in 14 gleich große Subkorpora mit einem Gesamtumfang von ca. 28 Millionen Wort-Token (vgl. Tabelle 1). Gemeinsame Datengrundlage sind die *Datenbank Gesprochenes Deutsch (DGD)* (Schmidt, 2017) und das *Deutsche Referenzkorpus (DeReKo)* (Kupietz et al., 2018) als jeweils umfassendste deutschsprachige Sammlung ihrer Art, sowie das Songkorpus (Schneider, 2020). Berücksichtigt werden ausschließlich komplette, ungekürzte Texte. Soweit mithilfe von Metadaten realisierbar, wurde bei der Stichprobenziehung auf eine diachron und regional ausgewogene Mischung geachtet.

Folgende Subkorpora enthalten **prototypisch konzeptionell mündliche Texte** aus den *DGD*-Gesprächskorpora:

(1) Deutsch Heute (DH): Regional stratifizierte gebrauchssprachliche Sammlung. Sie „umfasst alle Nationen und Regionen Mitteleuropas, in denen Deutsch heute als Amts- und Unterrichtssprache verbreitet ist“ (Kleiner, Berend, Knöbl & Brinckmann, 2014, S. 184). In unser Untersuchungskorpus fließen daraus nur Aufnahme-

[3] COAST (Conceptual Orality Analysis and Scoring Tool); online unter `https://github.com/rubcompling/COAST`.

transkripte freier Rede ein, zumeist biografische Erzählungen mit wechselnden Themenschwerpunkten. Vorleseaufgaben, die ebenfalls zum DH-Inventar zählen, werden explizit ausgefiltert.

(2) Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK): Transkribierte Audioaufzeichnungen authentischer Spontansprache, „die verschiedenste Interaktionstypen aus den Bereichen privater (z. B.Tischgespräche, Telefongespräche, Spielinteraktionen, Gespräche bei privaten Aktivitäten), institutioneller (z. B. schulischer Unterricht, Verkaufsgespräche, Fahrschulstunden, berufliche Gespräche, universitäre Prüfungsgespräche) und öffentlicher Kommunikation (z. B. Podiumsdiskussion, Schlichtungsgespräch) abdecken (Schmidt, 2018, S. 216)“.

Folgende Subkorpora enthalten **prototypisch konzeptionell schriftliche Texte** aus den DeReKo-Schriftsprachkorpora:

(3) Zeitschriften: Zeitlich (diachron) ausgewogenes Sample der Jahrgänge 1970 bis 2021 des Nachrichtenmagazins *Spiegel*. Angesichts der über diesen Zeitraum erwartbaren Autorenfluktuation und dem damit verbunden vernachlässigbaren Risko unregelmäßiger Häufungen („Clumpiness“) einzelner Autoren sowie der ebenfalls gegebenen Streuung der regionalen Autorenherkunft verzichten wir auf die Heranziehung weiterer Pressetitel. Zur Vermeidung möglicher Überschneidungen mit dem Subkorpus 'Interview' (s.u.) werden keine als 'Gespräch' oder 'Interview' deklarierten Beiträge einbezogen.[4]

(4) Zeitungen: Zufällige Auswahl von Inhalten des *Mannheimer Morgen*, zeitlich ausgewogen stratifiziert über die zurückliegenden vier Jahrzehnte. Analog zum Zeitschriftensubkorpus wird auf eine weitergehende Diversifizierung nach Pressetitel verzichtet.

Das Untersuchungskorpus weist damit eine weitestgehend ausgeglichene Menge an mutmaßlich konzeptionell schriftlichen Tokendaten und konzeptionell mündlichen Tokendaten auf (jeweils ca. 4 Millionen). Wohl wissend, dass ein einzelner Text grundsätzlich sowohl typische Nähe- als auch typische Distanzmerkmale aufweisen kann, lautet die Prämisse unserer Studie: Die Zuordnung der jeweils kompletten o.g. Texte zu einem der beiden Pole erlaubt das Trainieren statistischer Modelle für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit.

Darüber hinaus umfasst unser Untersuchungskorpus verschiedene ad hoc **nicht eindeutig zuordenbare Textsammlungen** mit ebenfalls je 2 Millionen Token, die sich mutmaßlich in unterschiedlichen Bereichen des Mündlich-Schriftlich-Kontinuums bewegen:

[4] Zur Veranschaulichung des quantitativen Hintergrunds: Sämtliche Jahrgänge der originalen *Spiegel*-Daten zusammengenommen umfassen ca. 210 Millionen Token, davon entfallen ca. 23,5 Millionen Token auf Gespräche/Interviews. In unser Untersuchungskorpus werden aus den verbleibenden Daten ca. 40.000 Token pro Jahr aufgenommen.

(5) Belletristik: Berücksichtigt wird Unterhaltungsliteratur (Romane und Erzählungen) des 20. und 21. Jahrhunderts. Das datengebende DeReko-Korpus enthält Texte diverser Schriftsteller im Umfang von knapp 10 Millionen Token, verfasst zwischen 1997 und 2012. Daraus werden für jedes Jahr komplette Texte mit insgesamt 2 Millionen Token zufällig extrahiert.

(6) DGDMISC: In dieses Subkorpus fließen weitere Aufnahmen gesprochener deutscher Sprache aus der *Datenbank Gesprochenes Deutsch* ein. Es enthält Auszüge zehn heterogener *DGD*-Korpora, z. B. 'Biographische und Reiseerzählungen', 'Berliner Wendekorpus', 'Dialogstrukturen', 'Elizitierte Konfliktgespräche', 'Gesprochene Wissenschaftssprache', König-Korpus, Pfeffer-Korpus.

(7) Interviews: Das Subkorpus speist sich aus drei DeReKo-Quellen: Zum einen enthält es *Stern*- und *Zeit*-Interviews aus den Jahren 1992-1994, weiterhin diverse zwischen 1999 und 2003 geführte politische Interviews und schließlich zeitlich gleichmäßig verteilte und zufällig ausgewählte *Spiegel*-Interviews 1970-2021.

(8) Liveticker: Berücksichtigt werden Live-Berichterstattungen aus der Domäne 'Fußball' zwischen 2006 und 2016. Das Subkorpus enthält zu gleichen Teilen „in Echtzeit geschriebene und im Internet publizierte Reportagen zu laufenden Spielen" (Meier-Vieracker, 2018, S. 3) aus den Korpora zur Fußballlinguistik sowie vom Fachportal *kicker.de*.

(9) Reden: Das Subkorpus versammelt zufällig ausgewählte und zeitlich ausgewogen verteilte Plenarprotokolle des Landtags von Rheinland-Pfalz aus den Jahren von 1996 bis 2012.

(10) SocialMedia: Abgedeckt werden interaktive Kommunikationskanäle in Form von Online-Kurznachrichten. Das Subkorpus enthält zu gleichen Teilen eine zufällige Auswahl 2021 aufgezeichneter Twitter-Tweets sowie Beiträge aus dem Dortmunder Chat-Korpus (Beißwenger, 2013) aus den Jahren 1998 bis 2006.

(11) SocialMedia-L: Das Subkorpus enthält längere Kurznachrichten zur Abschwächung kürzebedingter Effekte. Es basiert auf den selben Quellen wie die vorstehende Auswahl; im Unterschied dazu werden Texte allerdings nicht zufällig ausgewählt, sondern pro Jahr diejenigen mit der höchsten Tokenanzahl priorisiert. Die Anzahl der Einzeltexte reduziert sich dadurch ungefähr um den Faktor 5.

(12) Songtexte: Das Subkorpus enthält deutschsprachige Songtexte der zurückliegenden fünf Jahrzehnte, untergliedert in autorenspezifische, thematische sowie popularitätsbasierte Subarchive (Charts); siehe (Schneider, 2019a). Abgedeckt werden ost- und westdeutsche Künstler sowie unterschiedliche Genres wie Hiphop, Liedermacher, Pop und Rock (Schneider, 2019b).

(13) Wikipedia Nutzerdiskussionen (WikiDisk): Berücksichtigt werden die Online-Diskussionsseiten der deutschsprachigen Wikipedia-Community (Margaretha &

Lüngen, 2014). Abgedeckt wird der Zeitraum zwischen 2002 und 2017 mit vom Umfang her jeweils ähnlichen jährlichen, zufällig extrahierten Anteilen.

(14) Wissenschaftsschriftsprache (WissSchrift): Das Subkorpus enthält arbiträr ausgewählte Artikel aus *Gingko (Geschriebenes ingenieurwissenschaftliches Korpus)* (Schirrmeister, Rummel, Heine, Suppus & Mendoza Sánchez, 2021), die zwischen 2007 und 2016 in den Fachzeitschriften 'Automobiltechnische Zeitschrift (ATZ)' und 'Motortechnische Zeitschrift (MTZ)' erschienen sind.

Einzeltextvolumina der Subkorpora unterscheiden sich aus naheliegenden Gründen mitunter erheblich: So sind z. B. die 2 Millionen Token in der Belletristik-Sammlung mit durchschnittlich 57.887 Token pro Text bereits bei 36 Texten erreicht, im Social-Media-Subkorpus – das viele extrem kurze Tweets und Chats mit durchschnittlich nur 36 Token pro Text enthält – erst bei 55.041 Texten.

Subkorpus	Texte	Token	Token/Text
DH	667	2.003.377	3.004
FOLK	648	2.001.021	3.088
Zeitschrift	3.011	2.003.071	665
Zeitung	8.505	2.000.598	235
Belletristik	36	2.083.944	57.887
DGDMISC	1.061	2.000.630	1.886
Interviews	1.010	2.002.315	1.983
Liveticker	1.444	2.001.636	1.386
Reden	49	2.006.773	40.955
SocialMedia	55.041	2.000.012	36
SocialMedia-L	12.349	2.008.262	163
Songtexte	7.455	2.002.538	269
WikiDisk	4.051	2.000.159	494
WissSchrift	1.199	2.000.314	1668

Tabelle 1: Umfang und Untergliederung des Untersuchungskorpus. Die ersten vier Subkorpora dienen als Trainingsdaten für die Klassifikatoren.

Sämtliche Primärdaten sind angereichert um Lemmaangaben sowie Wortklassenannotationen nach dem für das Deutsche einschlägigen Stuttgart-Tübingen-Tagset (STTS). Die DGD-basierten Subkorpora verwenden das für gesprochene Sprache modifizierte STTS 2.0 (Westpfahl, Schmidt, Jonietz & Borlinghaus, 2017), Songtexte eine nochmals für spezielle Phänomene -– insbesondere zusammengezogene Wörter — erweiterte Spezifikation (Schneider, 2022, S. 45).

3.2 Das Featureset

Die von Ágel und Hennig (2006a, 2006b) konzipierte Systematik, die linguistisch fundiert eine Begründung des Nähestatus einzelner Merkmale liefert, lässt sich ohne tiefgehende computerlinguistische Analyse und Annotation der zur Disposition stehenden Textdaten schwerlich operationalisieren. Wir orientieren uns deshalb zunächst am von Ortmann und Dipper (2019) verwendeten Featureset und erweitern dieses um zusätzliche linguistisch motivierte Merkmale bzw. lassen diejenigen Features außen vor, die sich für das Untersuchungskorpus nicht berechnen lassen. Letzteres betrifft satzbasierte Merkmale wie Satzlänge, Satzeinleiter und Satztyp ('Fragesatz', 'Ausrufesatz'), weil verschriftete mündliche Daten zumeist ohne Interpunktionszeichen daherkommen. In der Konsequenz kann unsere Modellierung als eine Art Machbarkeitsstudie betrachtet werden, die opportunistisch zunächst mit einer Reihe verfügbarer Merkmale arbeitet in der Hoffnung auf empirisch fundierte Hinweise zu deren Relevanz und Gewichtung.[5]

In der nachfolgenden Übersicht unterscheiden wir Merkmale, die ausschließlich unter Rückgriff auf die Textoberfläche generiert werden (Tabelle 2), von solchen, die für die Identifizierung von Belegen ergänzend den Output (maschineller) Tagger heranziehen (Tabelle 3) und damit u.a. von deren Methodik, Annotationsinventar und Güte abhängen.

Durch das Heranziehen unterschiedlicher Maße mit variierenden methodischen Details zielen wir auf die Erarbeitung konvergierender Evidenz bei der optimalen konzeptuellen Einstufung (*Permutation Feature Importance* und *T-Test*) bzw. informationeller Mehrwerte hinsichtlich der Wirkungsrichtung (Δ):

Die *Permutation Feature Importance* gibt die Random Forest-Schätzung der mittleren Abnahme an Genauigkeit bei Weglassen eines Features wieder.

Wir prüfen mit dem *Welch-T-Test*, ob sich die Datenmittelwerte eines Features für beide Klassen (konzeptionell mündlich bzw. konzeptionell schriftlich) unterscheiden, mit einer Irrtumswahrscheinlichkeit von 0.05 (*), 0.01 (**) und 0.001 (***).

Δ charakterisiert in Vorzeichenform die Wirkungsrichtung eines Features, basierend auf der Differenz zwischen seinen Mittelwerten für konzeptionell mündliche vs. konzeptionell schriftliche Texte (+ für höheren Mittelwert bei mündlichen Texten bzw. − für höheren Mittelwert bei schriftlichen Texten).

3.2.1 Merkmale ohne Tagging-Informationen

Folgende **lexikalische Merkmale**, die sich ohne Zuhilfenahme computerlinguistischer Annotationen direkt auf der Textoberfläche bestimmen lassen, fließen in unsere Klassifizierung ein:

- Lexikalische Vielfalt wird bereits von Koch und Oesterreicher (1985, S. 454) als Unterscheidungskriterium angeführt und beruht auf der Annahme einer tendenziell höheren lexikalischen Varianz in Distanztexten. Hier wie bei allen statistischen Merkmalen gilt: Selbstverständlich sind Kommunikationssituationen (=

[5] Wir danken Mathilde Hennig für wertvolle Anregungen zur linguistischen Einordnung vieler dieser Merkmale.

Merkmal	Typ	Beschreibung	Δ	T-Test
STTR	lexikalisch	Lexikalische Vielfalt (Standardisierte Type-Token Ratio	–	***
MATTR	lexikalisch	Lexikalische Vielfalt (Moving-Average Type-Token Ratio)	–	***
MLTD	lexikalisch	Lexikalische Vielfalt (Measure of Textual Lexical Diversity)	–	***
PRON1st_wf	lexikalisch	Verhältnis Personalpronomina 1. Pers. zu allen Wörtern	+	***
PRON2nd_wf	lexikalisch	Verhältnis Personalpronomina 2. Pers. zu allen Wörtern	+	***
bloss	lexikalisch	Verhältnis Gradpartikel *bloß* zu allen Wörtern	+	***
lediglich	lexikalisch	Verhältnis Gradpartikel *lediglich* zu allen Wörtern	–	***
kriegen	lexikalisch	Verhältnis Verb *kriegen* zu allen Wörtern	+	***
word_mean	morphologisch	Wortlänge (Mittelwert)	–	***
word_med	morphologisch	Wortlänge (Median)	–	***
elision_ART	morphologisch	Verhältnis indefinite Artikel mit Elision zu Formen ohne Elision	+	***
elision_VA	morphologisch	Verhältnis Hilfsverben mit Elision zu Formen ohne Elision	+	***
VERBshort	morphologisch	Verhältnis gekürzter Verbformen nach *ich* zu allen Verbformen nach *ich*	+	***
contracted	morphologisch	Verhältnis Wortkontraktionen zu allen Wörtern	+	***
PTK_MOD	pragmatisch	Verhältnis Modalpartikeln zu allen Wörtern	+	***
repetition	pragmatisch	Verhältnis Wortwiederholungen zu allen Wörtern	+	***
stretch	pragmatisch	Verhältnis Wörter mit Buchstaben-Reduplikation zu allen Wörtern	–	**

Tabelle 2: Klassifikations-Features ohne Tagging-Informationen.

Texte) denkbar, in denen das Merkmal ohne Effekt bleibt oder sogar in eine andere Richtung weist. Anstelle einfacher Type-Token-Ratios (TTR), die stark mit Textgrößen korrelieren und deshalb wenig aussagekräftig scheinen, evaluieren wir drei differenziertere Maße: (a) *Standardized Type-Token Ratio (STTR)* zerteilt Texte in aufeinander folgende Fenster von 100 Wörtern und berechnet deren TTR-Gesamtmittelwert; (b) *Moving-Average Type-Token Ratio (MATTR)* von Covington und McFall (2010) arbeitet ebenfalls mit Wortfenstern, verschiebt diese allerdings sequenziell und begegnet damit dem Problem übrig bleibender Segmente mit weniger als 100 Wörtern; (c) *Measure of Textual Lexical Diversity (MLTD)* schließlich ist ein von Mccarthy und Jarvis (2010) beschriebenes komplexes Maß, das in mehreren Berechnungsschritten und -richtungen einen vergleichsweise robusten Vielfaltswert kalkuliert. Unvorteilhaft bleibt, dass sich keines dieser Maße uneingeschränkt für sehr kurze Texte empfiehlt.

- Mit den Maßen *PRON1st_wf* und *PRON2nd_wf* messen wir die Verwendung der ersten und zweiten Person in Texten und kalkulieren relative Frequenzen der Personalpronomen *ich, mich, mir, wir, uns* bzw. *du, dich, dir, ihr, euch.* Motiviert ist das Maß vom mutmaßlich größeren Bedarf in Nähetexten; vgl. u.a. Ágel/Hennig oder Biber, Johansson, Leech, Conrad und Finegan (1999).

- Stilistisch könnten Gebrauchsfrequenzen der Gradpartikeln *bloß* vs. *lediglich* auf einen der beiden Pole hindeuten: „Vor allem mündlich wird *bloß* vorgezogen, insbesondere in informellen Kontexten. *Lediglich* findet sich vorzugsweise in geschriebener Sprache“ (Zifonun et al., 1997, S. 877).

- Ähnliches gilt für die Verwendung von *kriegen* als Auxiliar des Dativpassivs; diese „gehört heute fast ausschließlich in die mündliche Umgangssprache“ (Zifonun et al., 1997, S. 1829). Wir mutmaßen eine insgesamt häufigere Verwendung der Wortform *kriegen* in Nähe/Mündlichkeit und messen die relativen Frequenzen in der Hoffnung auf klassifikatorischen Ertrag.

Weiterhin untersuchen wir **morphologische Merkmale** im Sinne von Features, die auf Phänomene der strukturellen Morphologie referieren:

- Längenverteilungen spielen eine prominente Rolle in der Sprachstatistik und bei der Formulierung quantitativer Sprachregularitäten. Wir betrachten Wortlängen in der Annahme, dass morphologisch komplexere - und damit längere – Wörter als elaborierter und damit tendenziell eher der Schriftsprache zuzuordnen sind. Berechnet werden Mittelwert (*word_mean*) sowie Median (*word_med*); zur Plausibilität und Verwendung für unser Sprachmodell vgl. Abschnitt 4.3.

- Elisionen als sprachökonomisch motivierte Vokalauslassungen im Wortinnern oder am Wortende treten „in mündlicher Sprache weit häufiger auf als in der – in dieser Hinsicht konservativeren – Schriftsprache“ (Grammis, 2020). Zur statistischen

Abbildung modellieren wir drei Merkmale: (a) *Elision_ART* berechnet das Verhältnis indefiniter Artikel mit Elision zu Formen ohne Elision, also z. B. *ne* bzw. *'ne* vs. *eine*; (b) *Elision_VA* steht für das Verhältnis von Hilfsverben mit Elision zu Realisierungen ohne Lautausfall unter Berücksichtigung sämtlicher Personen und Zeitformen also z. B. *hab, hab', hatt, hatt', werd, werd'* usw. vs. *habe, hatte, werde* usw. (c) *VERBshort* drückt das Verhältnis gekürzter Verbformen nach *ich* zu allen Verbformen nach *ich* aus. Berücksichtigt werden also nicht auf *-e* endende Realisierungen wie *ich geh* vs. *ich gehe*.[6]

- Zusammenziehungen aufeinander folgender Wörter gelten als ebenfalls charakteristisch für den mündlichen Diskurs und entsprechen wie auch Elisionen dem „auf das Verfahren der Sprecheinheitenbildung rückführbar[en]" Merkmal 'phonisches Wort' bei Ágel und Hennig (2006a, S. 60). Unser Merkmal *contracted* misst das Verhältnis von Wortkontraktionen – also *gibts, gibt's, haste, kannste, machste* usw. – jeweils zu allen Textwörtern.

Schließlich evaluieren wir **pragmatische Merkmale**, die Sprachdiskurse strukturieren, Einzelaspekte verstärken oder Stimmungen ausdrücken:

- Modal-/Abtönungspartikeln wie *auch, bloß, denn, doch, eben, eh* usw. „operieren auf Erwartungen und Einstellungen" der Kommunikationspartner und können gesprächssteuernd wirken, indem sie dazu beitragen, „Äußerungen in den jeweiligen Handlungszusammenhang zu integrieren" (Grammis, 2018). Ob sie sich damit für eine Verortung im Nähe-Distanz-Kontinuum eignen, überprüfen wir anhand des Merkmals *PTK_MOD*. Mangels passgenauer STTS-Annotation operieren wir dabei notgedrungen listenbasiert auf der Wortoberfläche, was eine Abgrenzung zu anderen Verwendungen verhindert und die Merkmalsschärfe mindert.

- Wortwiederholungen können als rhetorische Figur zur Aussageverstärkung eingesetzt werden; vgl. 'holistische Gefühlsäußerung durch Reduplikation' bei Ágel/Hennig. Mithilfe des Merkmals *repetition* berechnen wir die relativen Häufigkeiten von mehrfachen Wortnennungen wie in *immer immer wieder* oder *sehr, sehr gut*.[7]

- Reduplikationen einzelner Buchstaben zur Emulierung emotionaler Prosodie finden sich prominent in Verschriftlichungen konzeptioneller Mündlichkeit; vgl. (Rehm, 2002), (Schneider, 2022). Unser Merkmal *stretch* misst relative Häufigkeiten solcher 'Stretchwörter' wie *soooo, suuuper, tschüßiiiii* oder *Jetzt geht's lo-oos.*

[6] Eichinger (2017, S. 312) weist darauf hin, „dass hier im Modus des Sprechens Formen erscheinen, die wie eine Abschleifung der geschriebenen Form durch die (normale) Geschwindigkeit des Sprechens (Allegro-Sprechen) erscheinen, es aber nicht sind oder zumindest nicht in jedem Fall sein müssen".

[7] Nicht berücksichtigt werden Wiederholungen von Wortfolgen wie *O Gott, o Gott.*

3.2.2 Merkmale mit Tagging-Informationen

Einige der vorstehend genannten Features lassen sich alternativ zur Erkennung auf der Sprachoberfläche auch (und ggf. sogar exhaustiver bzw. exakter) unter Zuhilfenahme computerlinguistischer Annotationen identifizieren. Andere sind zwingend auf Angaben zu Wortklasse (POS) oder Lemma angewiesen. In unsere Untersuchung fließen deshalb auch Merkmale ein, die auf solchen maschinell erstellten Tagging-Informationen basieren.

Dies betrifft folgende **lexikalische Merkmale**:

- Bereits Halliday (1985) nimmt eine geringere lexikalische Dichte in Gesprochenem an. Für einzelne Texte berechnen wir dafür das Merkmal *lexDens* als Verhältnis von Inhaltswörtern (Autosemantika: Nomina, Vollverben, Adjektive, Adverbien) zu allen Textwörtern.

- Demonstrativpronomina können im Diskurs eingesetzt werden, um auf Personen, Gegenstände oder Sachverhalte im Wahrnehmungsfeld zu verweisen. Eine durch verstärkten Gebrauch eventuell verbundene Hinweisfunktion auf Nähesprache kodiert das Merkmal *DEM* und nutzt für die Erkennung das STTS-Tag *PDS*. Flankierend berechnen wir als Feature *DEMshort* das Verhältnis kurzer Demonstrativpronomina wie *der, die, das, den, dem* usw. zu allen Demonstrativpronomina, also incl. der Langformen *dies, diese, dieser, diesen, dieses*; zur Unterscheidung ziehen wir zusätzlich die Lemmata heran (*die, d* vs. *diese, dies, die, d*).

- *PRON1st* und *PRON2nd* kodieren analog zu *PRON1st_wf* und *PRON2nd_wf* relative Frequenzen der Personalpronomina. Die Vorkommen ermitteln wir hier auf Basis der POS-Tags (*PPER*) sowie der Lemmata (*ich, wir* bzw. *du, ihr*).

- Das Merkmal *V_N* dient einer statistischen Annäherung an Verbal- bzw. Nominalstil. Ersterer gilt als lebendiger und eher umgangssprachlich anzutreffen, letzterer findet sich eher in sprachökonomisch optimierten (fachlichen) Distanztexten. Stark vereinfachend untersuchen wir hierfür das Verhältnis von Vollverben zu Nomina in den einzelnen Texten.

Ergänzende **morphologische Merkmale** mit Tagging-Informationen beschränken sich auf die konkatenative Morphologie ausgewählter Wortklassen:

- Die Merkmale *autosem_mean* bzw. *autosem_med* bestimmen Mittelwert bzw. Median der Wortlänge sämtlicher Autosemantika eines Texts. Auf diese Weise konzentrieren wir uns auf Wortklassen, die morphologische Komplexität durch Komposition oder Flexion überhaupt erst ermöglichen. Synsemantika – meist ohne Längenvariation – beeinflussen hierbei nicht den statistischen Aussagewert.

Als **pragmatische Merkmale** im Kommunikationszusammenhang kommen hinzu:

- Interjektionen als Merkmale mündlicher Kommunikation können spontane Empfindungen ausdrücken (*ächz, au, seufz* usw.), als Pausen-/Überbrückungs- (*äh,*

hm usw.) oder Aufmerksamkeitsmarker (*hey, ey* usw.) dienen und vieles mehr. Für die Berechnung des Features *INTERJ* beschränken wir uns auf Ein-Wort-Interjektionen.

- Antwortpartikeln sind typische hörerseitige Diskursbeiträge. Für das Merkmal *PTKANT* filtern wir nicht nach Wortform oder Lemma, sondern nutzen die STTS-Wortklassenannotationen. Neben *ja* und *nein* werden auch Formen wie *jaaa, nee, gewiss, bitte, danke* etc. abgedeckt.

Auch wenn sich womöglich aussagekräftigere Merkmale wie fehlerhafter Satzbau oder Selbstkorrekturen ohne tiefergehendes syntaktisches Tagging nicht algorithmisiert identifizieren lassen, können basierend auf den vorhandenen Annotationsebenen doch einige **syntaktische Merkmale** bestimmt werden:

- Passivstil deutet tendenziell auf Konstruiertheit und damit Distanzverortung eines Texts hin. Für das Feature *passiv* zählen wir – massiv vereinfacht und als experimentelle Annäherung – auf der Oberfläche erkennbare *werden*-Passive und berücksichtigen hierfür Partizip II-Verbformen unmittelbar nach dem passivbildenden Hilfsverb (realisiert als *werde, wirst, wird, werden, werdet, wurde, wurdest, wurden, wurdet*) bzw. mit maximal einem Zwischenwort; auch Belege mit umgekehrter Anordnung ohne optionale Lücke fließen in die Berechnung ein: *Der Bürgermeister wurde gestern abgewählt* oder *Das Problem soll geklärt werden.*

- Spontansprache gilt als syntaktisch weniger komplex als geplante Sprache. Das Feature *subord* dient der Identifizierung von Hypotaxen, also der Unterordnung von Nebensätzen in Satzgefügen. Wir setzen hierzu – wiederum stark vereinfachend – vom Tagger als unterordnende Konjunktionen erkannte Vorkommen von *um, anstatt, weil, dass* usw. in Beziehung zu allen Vollverben.

4 Textklassifikation

Wir trainieren binäre Klassifikatoren zur Unterscheidung konzeptioneller Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit auf Einzeltextebene. Neben der Klassifikationsgüte bewerten wir die Relevanz der oben eingeführten Merkmale. Anschließend verorten wir auch nicht-eindeutige Texte im Kontinuum.

Hierzu werden verschiedene Klassifikatoren auf einem Teil der vier in Tabelle 1 erstgenannten Subkorpora trainiert (Abschnitt 4.1). Für DH und FOLK wird dabei konzeptionelle Mündlichkeit und für die beiden DeReKo-Subkorpora (Zeitschriften und Zeitungen) konzeptionelle Schriftlichkeit angenommen; jeder Einzeltext ist ein Datenpunkt. Damit führen wir die statistischen Berechnungen auf Basis authentischer, breit gestreuter und gleichzeitig zeitgenössischer Datensamples durch, um die Modellierung der beiden Pole auf ein solides Fundament zu stellen. In Abschnitt 4.3 ermitteln wir redundante Variablen, die dann für die nachfolgenden Schritte ausgeschlossen werden.

Merkmal	Typ	Beschreibung	Tagging	Δ	T-Test
lexDens	lexikalisch	Verhältnis Autosemantika (ADJ.*, ADV, N.*, VV.*) zu allen Wörtern	POS	+	***
DEM	lexikalisch	Verhältnis Demonstrativpronomina (PDS) zu allen Wörtern	POS	+	***
DEMshort	lexikalisch	Verhältnis PDS-Kurzformen zu allen Demonstrativpronomina	POS, Lemma	+	***
PRON1st	lexikalisch	Verhältnis Personalpronomina (PPER) 1. Pers. zu allen Wörtern	POS, Lemma	+	***
PRON2nd	lexikalisch	Verhältnis Personalpronomina (PPER) 2. Pers. zu allen Wörtern	POS, Lemma	+	***
V_N	lexikalisch	Verhältnis Vollverben zu Nomina	POS	+	***
autosem_mean	morphologisch	Länge Autosemantika (Mittelwert)	POS	–	***
autosem_med	morphologisch	Länge Autosemantika (Median)	POS	–	***
INTERJ	pragmatisch	Verhältnis Ein-Wort-Interjektionen zu allen Wörtern	POS	–	***
PTK_ANT	pragmatisch	Verhältnis Antwortpartikeln zu allen Wörtern	POS	–	***
passiv	syntaktisch	Verhältnis partizipiale Verbformen vor oder hinter *werden*-Hilfsverb zu allen Verbformen	POS	–	***
subord	syntaktisch	Verhältnis unterordnender Konjunktionen (KOUS, KOUI) zu Vollverben	POS	+	***

Tabelle 3: Klassifikations-Features mit Tagging-Informationen.

Auf dieser Basis trainieren und evaluieren wir in Abschnitt 4.4 Klassifikatoren mit verschiedenen Feature-Teilmengen. Zwei dieser Klassifikatoren setzen wir schließlich in Abschnitt 4.5 für die Einordnung von Texten ein, die – wie z. B. Songtexte – intuitiv nicht eindeutig einer polaren Konzeption zugeordnet werden können und mutmaßlich irgendwo im Kontinuum zwischen Oralität und Literalität liegen.

4.1 Einteilung der Datensets

Die durchschnittlichen Primärtextgrößen unterscheiden sich zwischen DGD- und DeReKo-basierten Subkorpora um den Faktor 5 bis 13, so dass trotz gleicher Gesamttokenzahl grundsätzlich wesentlich mehr konzeptionell schriftliche Texte als konzeptionell mündliche Texte für das Training des Klassifikators zur Verfügung stehen. Da das Trainieren auf eine hohe generelle Genauigkeit (*accuracy*) abzielt, besteht hier die Gefahr, dass die Accuracy für die Minderheitenklasse (hier *specificity*) wesentlich schlechter ist als die für die Mehrheitsklasse (hier die *sensitivity*). Dem begegnen wir durch partielle Reduktion der Trainingsdaten (*downsampling*): Es werden aus der Menge der konzeptionell schriftlichen Texte zufällig die gleiche Anzahl an Texten gewählt, wie sie maximal für konzeptionelle Mündlichkeit bereitstehen.[8].

Konzeption	Texte
schriftlich	11516
mündlich	1315

Tabelle 4: Verteilung der Texte nach konzeptioneller Verortung bei gleicher Tokenzahl.

Die verbleibenden Daten werden im Verhältnis von 80% zu 20% in ein Development- und ein Validierungsset aufgeteilt. Das Developmentset teilen wir wiederum im 80/20-Verhältnis in ein Trainings- und ein Testset (*Out-of-Bag*-Stichprobe (OOB) zur unvoreingenommenen Bewertung z.B. von Parameteranpassungen beim Trainieren der Modelle). Tabelle 5 verdeutlicht die Aufteilung.[9]

4.2 Software und statistischer Ansatz

Die vorliegende Studie implementiert ein überwachtes Lernverfahren („supervised learning“) in Form eines Random-Forest-Algorithmus. Random Forest ist eine nichtparametrischer Klassifikationsmethode, die Ergebnisse verschiedener – zufällig und unkorreliert generierter – Entscheidungsbäume („decision trees“) kombiniert und ihre

[8] *Upsampling/bootstrap sampling* der Minderheitenklasse ist zwar grundsätzlich auch möglich, aber der Einfluss der unabhängigen Variablen ließe sich dann nicht zuverlässig bestimmen (Strobl, Boulesteix, Zeileis & Hothorn, 2007)

[9] Da die Gütekriterien auf den Trainingsdaten gemeinhin immer gut sind, wird die Güte auf Grundlage des Testsets bewertet und der Klassifikator entsprechend entwickelt. Um ein Overfitting auf das Testset zu vermeiden, wird nach Abschluss der Entwicklung die Güte des fertigen Klassifikators mit dem Validierungsset ermittelt.

Konzeption	Trainingsset	Testset	Validierungsset	
schriftlich	839	213	263	1.315
mündlich	844	208	263	1.315
	1.683	421	526	2.630

Tabelle 5: Anzahl der Texte nach Konzeptionsklassen pro Datenset.

Zuordnungen auf Basis dieser Einzelentscheidungen vornimmt. Das Verfahren wird als Stichprobennahme ohne Zurücklegen („subsampling without replacement“) realisiert, weil sich damit der Effekt konfundierender Variablen auf das Importance Measure verringert: Unser Vorgehen vermeidet, dass einige uninformative Variablen aufgrund ihrer Struktur häufiger einfließen als andere (vgl. Strobl et al., 2007).

Wir nutzen R in der Version 4.2.0 (R Core Team, 2022) und *cforest* aus dem *party*-Package (Hothorn, Buehlmann, Dudoit, Molinaro & van der Laan, 2006; Strobl, Boulesteix, Kneib, Thomas & Zeileis, 2008; Strobl et al., 2007) mit Default-Parametern. Abbildungen und Tabellen sind mit *ggplot2* (Wickham, 2016) und unter Zuhilfenahme von *xtable* (Dahl, Scott, Roosen, Magnusson & Swinton, 2019) erstellt. Für die *Variable Importance Measures* (VIMs) kommt das Package *permimp* (Debeer, Hothorn & Strobl, 2021) zum Einsatz. Die Güte-Werte wurden mit Hilfe von *caret* (Kuhn, 2022) extrahiert.

4.3 Ausschluss redundanter Variablen

Tokenlängen liegen als Durchschnittswerte (`word_mean` / `autosem_mean`) und als Mediane (`word_median` / `autosem_mean`) vor. Mediane sind robuster bei Ausreißern, allerdings gibt es Informationsverluste in Bezug auf Extrempunkte. Durchschnittswerte sind datengetreuer, andererseits können hier Ausreißer den Wert verzerren. Eine Klassifizierung mit beiden Längenmaßen erscheint unnötig. Um zu entscheiden, welche Variablen final Verwendung finden sollen, trainieren wir einen ersten Random Forest mit allen Features.

Auf Basis der *Permutation Importance* (*Mean Decrease Accuracy*, siehe Abbildung 1) zeigen sich dabei die Durchschnittswerte (`autosem_mean` und `word_mean`) als aussagekräftiger. Die Importance bleibt auch bei bei mehrmaligem Trainieren stabil.

Bestätigend beobachten wir, dass die Güte eines Klassifikators mit ausschließlich diesen beiden Prädiktoren am besten ist (Tabelle 6) und es eine Verbesserung sowohl zu Klassifikatoren mit jeweils nur einem Prädiktor als auch zu solchen mit anderen Variablenkombinationen gibt. Die Unterschiede sind allerdings nicht groß und betragen meist $< 1\,\%$, mit Ausnahme eines Klassifikators (`word_med`), der ca. 6,4 % weniger Testset-Daten richtig und ca. 13,7 % der konzeptionell mündlichen Daten falsch klassifiziert.

4.4 Vergleich verschiedenartiger Klassifikatoren

4.4.1 Einbezogene Features

Insgesamt trainieren wir Modelle für drei Klassifikatoren:

- Ein Klassifikator K_{Gesamt}, der – unter Ausschluss redundanter Features (s. Unterabschnitt 4.3) – alle Variablen beinhaltet.
- Ein Klassifikator $K_{Tagging}$ nur mit Variablen, die auf Tagging-Informationen basieren (s. Tabelle 3). Fehlannotationen können sich dabei auf die Performanz des Klassifikators auswirken.[10]
- Ein Klassifikator $K_{Oberfläche}$ nur mit Variablen, für deren Berechnung keine Tagging-Informationen genutzt werden, um den Einfluss eventueller Annotationsfehler abzuschätzen.

UV	Acc.	Acc.P.	B.Acc.	F1	Prec.	Sens.	Spec.
word_mean	0,9810	< 0,001	0,9810	0,9812	0,9812	0,9812	0,9808
word_med	0,9216	< 0,001	0,9315	0,9160	0,8451	1,0000	0,8631
autosem_mean	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_mean + word_mean	0,9857	< 0,001	0,9858	0,9860	0,9906	0,9814	0,9903
autosem_mean + word_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med + word_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med + word_mean	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855

Tabelle 6: OOB-Güte für Klassifikatoren mit permutierten `autosem-` und `word-` Features.

4.4.2 Klassifikationsgüte der polaren Texte

Tabelle 7 illustriert, dass die Daten im Validierungsset beinahe ausnahmslos korrekt eingestuft werden. Am besten schneidet $K_{Oberfläche}$ ab, der Unterschied zu den beiden anderen Klassifikatoren ist jedoch kaum spürbar (0,0038 Accuracy-Punkte). Hier könnte sich der Umstand auswirken, dass ohne Rückgriff auf maschinelle Annotationen berechnete Variablen störunanfälliger sind.

[10] Der auf standardnahe Sprache trainierte Tagger weist Schwächen etwa bei der Beurteilung von Wortzusammenziehungen, Interjektionen und Named Entities in 'standardferneren' Texten auf.

Klassifikator	Acc.	Acc.P.	B.Acc.	F1	Prec.	Sens.	Spec.
Gesamt	0,99429	$< 0{,}001$	0,9943	0,9943	0,9962	0,9924	0,99618
Tagging	0,99429	$< 0{,}001$	0,9943	0,9943	0,9962	0,9924	0,99618
Oberfläche	0,99809	$< 0{,}001$	0,9981	0,99809	0,9962	1,0000	0,99621

Tabelle 7: Güte der Klassifikatoren auf dem Validierungsset.

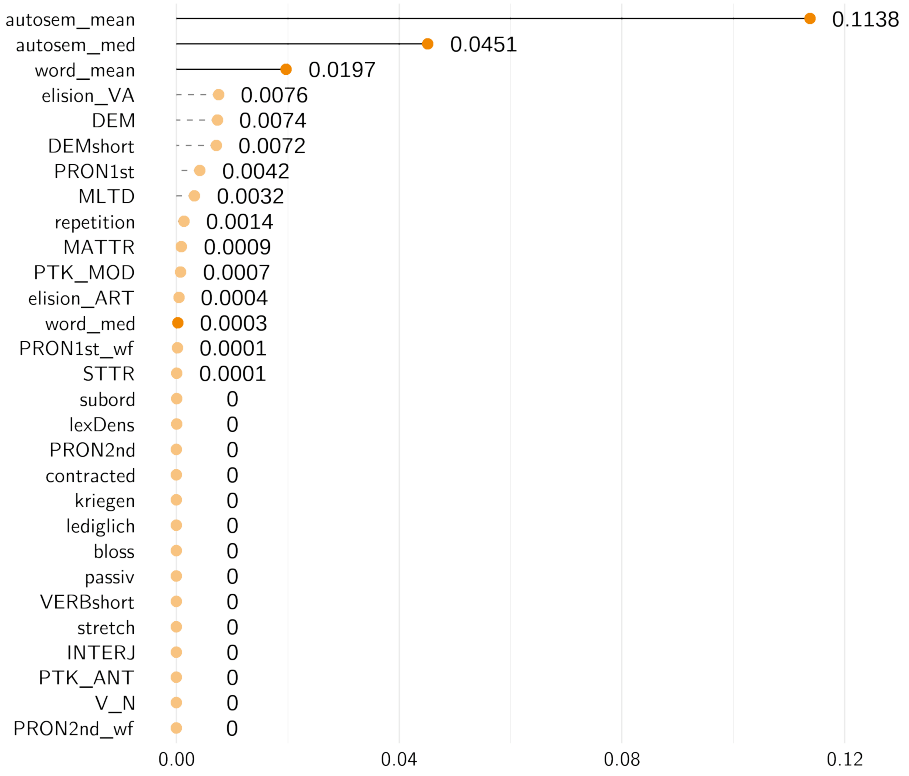


Abbildung 1: Permutation Importance für alle untersuchten unabhängigen Variablen. Die farblich hervorgehobenen Variablen sind untereinander redundant.

Die Accuracy (Acc.) ist bei allen Klassifikatoren hochsignifikant besser als die *No Information Rate*, also die beste Schätzung, wenn keine über die Gesamtverteilung der Klassen hinausgehenden Informationen vorliegen (Acc.P. mit $p < 0{,}001$). Der Accuracy-Durchschnitt (*Balanced Accuracy*, B.Acc.) fällt durchgehend nochmals minimal besser aus; dieser Unterschied lässt sich darauf zurückführen, dass die Anzahlen der Texte mit schriftlicher oder mündlicher Konzeption im Validierungsset nicht perfekt ausgeglichen sind (vgl. Tabelle 5).

Die F1-Werte, das harmonische Mittel aus *Precision* (der Anteil der tatsächlich schriftlichen Texte an allen als schriftlich klassifizierten Texte) und *Recall* (gleich *Sensitivity*, der Anteil aller richtig als schriftlich klassifizierten Texte an allen tatsächlich schriftlichen Texte unabhängig von ihrer Klassifikation), unterscheiden sich ebenfalls nur minimal von Acuracy bzw. Balanced Acuracy.

Die Accuracy-Werte legen nahe, dass alle drei Klassifikatoren zuverlässig arbeiten. Die Modelle unterscheiden sich nur geringfügig darin, wie zuverlässig sie Texte sprachlicher Konzeption (*Sensitivity* und *Precicion*) und Texte mündlicher Konzeption richtig klassifizieren (*Specificity*). Die *Precision* (Prec.) fällt bei allen Klassifikatoren gleich aus, jedoch ist sie bei K_{Gesamt} und $K_{Tagging}$ besser als die *Sensitivity* (Sens.). Bei $K_{Oberfläche}$ dagegen ist die *Sensitivity* besser als die *Precision*.

Die *Specificity* ist bei K_{Gesamt} und $K_{Tagging}$ gleich, bei $K_{Oberfläche}$ einen Hauch besser.[11] Von $K_{Oberfläche}$ werden Texte schriftlicher Konzeption gar nicht (*Sensitivity* = 1) und Texte mündlicher Konzeption (minimal) seltener fehlklassifiziert (*Specificity*).

4.5 Verortung der nicht-polaren Texte

Wir nutzen K_{Gesamt} und den insgesamt leistungsstärksten Klassifikator $K_{Oberfläche}$ für die experimentelle Verortung konzeptioneller Mischfälle auf Textebene – also derjenigen Texte, die wir nicht einem der beiden Pole zuschlagen und deren Einordnung im Kontinuum insofern ungesichert ist. Außer den 10 bislang nicht klassifizierten Subkorpora aus Tabelle 1 beziehen wir auch die beim Downsampling herausgefallenen schriftsprachlichen Texte ein. Dabei betrachten wir, ob die oben beschriebenen minimalen Güteunterschiede einen wahrnehmbaren Einfluss auf die Klassifikation nehmen.

Tabelle 8 ordnet die Subkorpora absteigend anhand des Anteils der von K_{Gesamt} als konzeptionell schriftlich eingestuften Texte. Wie erwartbar liegen die (zusätzlichen) Zeitungs- und Zeitschriftentexte mit über 99% am oberen Ende der Skala, werden allerdings noch übertroffen durch die wissenschaftssprachlichen Texte (ebenfalls intuitiv erwartbar) und Reden (nicht überraschend in Anbetracht der für Plenarprotokolle üblichen Vor- bzw. Nachbearbeitung). Die Klassifizierung der Zeitungs- und Zeitschriftentexte durch $K_{Oberfläche}$ fällt geringfügig eindeutiger aus, was sich mit der besseren *Sensitivity* (siehe Tabelle 7) des Modells deckt. Ein Unterschied der Sensitivität um nur 0,0076 Punkte (0,76 Prozentpunkte) macht hier demnach einen Accuracy-Unterschied von 0,2 % (Zeitungen) bzw. 0,5 % (Zeitschriften) aus.

[11] Dieser Unterschied hat bei der Klassifikation von konzeptionell mündlichen Texten einen Einfluss, wie in Abschnitt 4.5 deutlich wird.

	Gesamt			Oberfläche		
	schriftlich	mündlich	%	schriftlich	mündlich	%
WissSchrift	1.199	0	100 %	1.199	0	100 %
Rede	49	0	100 %	49	0	100 %
Zeitung	7.499	25	99,7 %	7.513	11	99,9 %
Zeitschrift	2.662	15	99,4 %	2.675	2	99,9 %
Liveticker	1.403	41	97,2 %	1.444	0	100 %
WikiDisk	3.846	205	94,9 %	4.001	50	98,8 %
Interview	955	55	94,6 %	978	32	96,8 %
SocialMedia	51.909	3.125	94,3 %	51.045	3.989	92,8 %
SocialMedia-L	10.773	1.576	87,2 %	11.469	880	92,9 %
Belletristik	28	8	77,8 %	31	5	86,1 %
Songs	3.965	3.490	53,2 %	2.255	5.200	30,2 %
DGDMISC	324	737	30,5 %	303	758	28,6 %

Tabelle 8: Textklassifikation mit Gesamt-Modell und Oberflächen-Modell. Die %-Spalten geben die prozentualen Anteile der als konzeptionell schriftlich eingestuften Texte an.

Inhalte des Subkorpus `DGDMISC` werden - ebenfalls erwartbar - als vorrangig mündlich klassifiziert, auch hier arbeitet $K_{Oberfläche}$ etwas eindeutiger als K_{Gesamt}. Der Unterschied in der *Specificity* um nur 0,003 Prozentpunkte macht hier einen 1,9-prozentigen Unterschied bei der Zuordnung aus.

Bemerkenswert sind die Klassifikationsergebnisse für einige andere Subkorpora, z.B.:

- Bei Interviews, also verschriftlichten Gesprächen, würde man intuitiv von einer tendenziell mündliche(re)n Konzeption ausgehen. Beide Klassifikatoren verorten allerdings den ganz überwiegenden Großteil dieser Texte als konzeptionell schriftlich.

- Der Einfluss der unterschiedlichen Textlängen in den beiden SocialMedia-Subkorpora fällt bei K_{Gesamt} größer aus als bei $K_{Oberfläche}$: Zweiterer klassifiziert die Inhalte beider Korpora fast gleich häufig als konzeptionell schriftlich bzw. mündlich, ersterer klassifiziert weniger SocialMedia-L-Texte (87,2 %) als SocialMedia-Texte (94,3 %) als konzeptionell schriftlich.

- Die beiden Modelle beurteilen Songtexte in Teilen unterschiedlich. K_{Gesamt} verortet die Datenpunkte (= Texte) weitestgehend ausgeglichen (3490 mündlich, 3965 schriftlich). $K_{Oberfläche}$ klassifiziert Songtexte mehrheitlich (69,8 %) als konzeptionell mündlich.

5 Diskussion

In diesem Abschnitt zeigen wir Gründe für unterschiedliche Klassifikationen auf und beleuchten überraschende Zuordnungen im Detail.

5.1 Falsche Zuordnungen im Validierungsset

Zunächst erörtern wir aus qualitativer Perspektive verschiedene Fehlklassifikationen im Validierungsset. Damit soll ein Gefühl für die Wirkungsweise der Klassifikatoren vermittelt sowie ein möglicher qualitativer Analyseansatz demonstriert werden.

Die Klassifikatoren haben im Validierungsset nur sehr wenige Texte nicht wie erwartet eingeordnet (vgl. Tabelle 7): K_{Gesamt} 3 von 526, $K_{Oberfläche}$ 1 von 526. Wir identifizieren folgende Faktoren, die Fehlklassifikationen begünstigen:

- Es gibt zum einen unter den einflussreichen Features mit den größten PI-Werten (Permutation Importance) Fälle, in denen der fehlklassifizierte Text innerhalb einer Spanne liegt, in der sich Werte konzeptionell mündlicher und schriftlicher Texte überschneiden.
- Zum anderen kann ein fehlklassifizierter Text innerhalb der Zielgruppe, der er eigentlich zugeordnet werden sollte, in Bezug auf ein (einflussreiches) Merkmal ein statistischer Ausreißer (Outlier) sein.
- Gleichzeitig kann er allerdings auch innerhalb der anderen Gruppe in Bezug auf das gleiche Merkmal in die Quartilsspanne fallen (wie beispielsweise Text 75645 bei `elision_VA`, der als Outlier unter den konzeptionell schriftlichen Texten in die Interquartilsspanne der konzeptionell mündlichen Texte fällt und vom gemischten Klassifikator fälschlicherweise als konzeptionell mündlich klassifiziert wird; siehe Abbildung 2).

5.1.1 Fehlklassifikationen des Gesamtmodells

K_{Gesamt} stuft Text 75645[12] fälschlicherweise als konzeptionell mündlich und 00361 bzw. 00159 fälschlicherweise als konzeptionell schriftlich ein. $K_{Oberfläche}$ schlägt Text 00260 unerwartet mündlicher Nähe zu. Nachfolgend inspizieren wir diese Texte genauer.

Ein Blick in die Primärinhalte enthüllt Text 75645 als kurzen Feuilleton-Zeitungsartikel mit popkulturellem Bezug. Darin finden sich keine komplexen Wörter und die durchschnittliche Wortlänge fällt entsprechend kurz aus. Dafür erkennt der Klassifikator fälschlicherweise eine auf Mündlichkeit hindeutende Hilfsverbelision (tatsächlich handelt es sich um die englische Wortform 'is') sowie mehrere mit Tagging-Unterstützung identifizierte kurze Demonstrativpronomina (Feature `DEMshort`). Der Oberflächen-Klassifikator plädiert auf konzeptionelle Distanz/Schriftlichkeit.

[12] Die originalen Text-Siglen sind hier aus Platzgründen gekürzt.

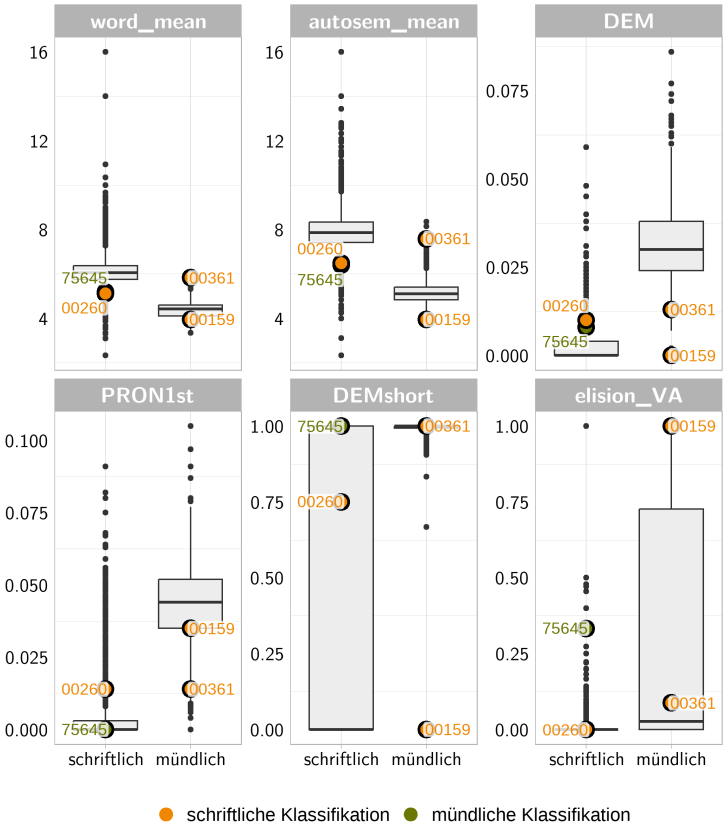


Abbildung 2: Boxplots für die sechs einflussreichsten Features des Gesamtmodells K_{Gesamt} mit farblich markierten Fehlklassifikationen.

FOLK-Transkript 00361 gibt eine längere wissenschaftliche Podiumsdiskussion wieder, 00159 die knappe Anmeldung an der Pforte eines Polizeireviers; beide Texte enthalten vergleichsweise wenig typische Nähe-/Mündlichkeitsmerkmale. Auch hier beurteilt $K_{Oberfläche}$ im Gegensatz zu K_{Gesamt} korrekt.

Abbildung 2 visualisiert die Situation. Sie zeigt die sechs für die Klassifikation maßgeblichsten Merkmale, d.h. Features mit der höchsten PI (vgl. dazu Abbildung 1).

Text 75645 liegt bei den einflussreichen Wortlängen-Features am unteren Ende des ersten Quartils für konzeptionell schriftliche Texte. Es besteht bei beiden Features eine Überschneidung des ersten Quartils für konzeptionell schriftliche Texte mit dem vierten Quartil der konzeptionell mündlichen Texte – und der fehlklassifizierte Text fällt genau in diese Überschneidung. Bei Merkmal `DEM` liegt der Text in der Überschneidung des vierten Quartils der konzeptionell schriftlichen Texte und des ersten Quartils der konzeptionell mündlichen Texte.

Wie aber an der Position des von K_{Gesamt} korrekt klassifizierten Zeitschriften-Texts 00260 deutlich wird, kann diese Ambiguität nicht der alleinige Grund für die Fehlklassifikation sein. Die Werte von `elision_VA` und `DEMshort` geben Aufschluss: Bei Merkmal `elision_VA` ist der Text ein potenzieller Outlier unter den konzeptionell schriftlichen Texten, liegt aber noch bequem im dritten Quartil der konzeptionell mündlichen Texte. Ebenso liegt der Text bei Merkmal `DEMshort` im Vergleich zu allen konzeptionell schriftlichen Texte am Maximum – wobei der Mittelwert beim Minimum liegt –, im Vergleich zu konzeptionell mündlichen Texten jedoch beim Mittelwert.

Im Gegensatz zu den Wortlängen-Werten von Text 75645 sind diese für den fälschlicherweise als konzeptionell schriftlich eingeordneten Text 00361 eindeutiger, denn der Text ist für die mündlichen Nähe-Texte ein Outlier, fällt jedoch ungefähr auf den Mittelwert der schriftlichen Distanz-Texte. Schließlich hat der ebenfalls als konzeptionell schriftlich klassifizierte Text 00159 einen für mündliche Nähe-Texte kleinen Anteil an Demonstrativpronomina und verkürzten Demonstrativpronomina (`DEM` und `DEMshort`) und ist damit ein potentieller Outlier; für konzeptionelle Distanz/Schriftlichkeit sind die Werte dagegen durchschnittlich.

5.1.2 Fehlklassifikation des Oberflächenmodells

Der von $K_{Oberfläche}$ unerwartet als Nähetext eingestufte Zeitschriftenartikel 00260 enthält bei näherer Betrachtung tatsächlich mehrere auf Mündlichkeit hindeutende Merkmale. Es handelt sich um eine mit direkter Rede angereicherte Buchbesprechung.

Für die Zuordnung scheinen insbesondere die Features `word_mean`, `PTK_MOD` und `PRON1st_wf` verantwortlich: Bei `word_mean` liegt der Artikel in der Überschneidung des ersten Quartils der konzeptionell schriftlichen Distanztexte und des vierten Quartils der konzeptionell mündlichen Nähetexte. Bei `PTK_MOD` fällt eine Überschneidung des vierten Quartils der konzeptionell schriftlichen Distanztexte und des ersten Quartils der konzeptionell mündlichen Nähetexte auf. Darüber hinaus ist der Artikel bei `PRON1st_wf` ein potenzieller Outlier unter den konzeptionell schriftlichen Distanztexten, nicht aber bei konzeptionell mündlichen Nähetexten.

Damit ist speziell bei den drei Oberflächen-Features mit der höchsten Permutation Importance (PI; vgl. Abbildung 4) eine gewisse Ambiguität gegeben.

5.2 Geringer Einfluss einzelner Merkmale

Einige Features haben einen sehr geringen oder gar keinen Einfluss auf die Textklassifikation. Deutlich wird das an einer niedrigen bzw. auf 0 lautenden Permutation Importance (PI; vgl. Abbildung 3). Im Falle der *Standardized Type-Token Ratio (STTR)* lässt sich konstatieren, dass die konkurrierenden und für unterschiedliche Text- und Korpusgößen robusteren Wortschatz-Maße *MATTR* bzw. *MLTD* erfreulicherweise einen höheren Einfluss aufweisen und damit die lexikalische Vielfalt doch recht prominent im Modell abgebildet wird. In anderen Fällen lässt sich eine plausibler Nullwerte-Zusammenhang konstatieren: So werden beispielsweise in über 80 % aller Texte keine Passivkonstruktionen erkannt – vermutlich nicht zuletzt, weil nur sehr vereinfacht auf der Oberfläche analysiert wurde. In solchen Fällen kann der Klassifikator dann auch keine Grenze für die Aufteilung der Daten anhand des Passiv-Merkmals festlegen.

Subkorpus	*kriegen*	*bloß*	*lediglich*
FOLK	46,3%	8,02%	0%
DH	27,74%	10,5%	0,15%
DGDMISC	16,21%	11,5%	2,5%
Songs	11,59%	5,7%	0,32%
Interview	9,2%	5,54%	4%
Liveticker	3,19%	0,42%	6,44%
Zeitschrift	2,82	2,42%	4,64%
Email-L	2,22%	1,84%	1,86%
SocialMedia-L	0,83%	0,28%	0,4%
Email	0,6%	0,55%	0,4%
Zeitung	0,47%	0,36%	2,7%
WikiDisk	0,44%	0,17%	1,28%
SocialMedia	0,27%	0,09%	0,05%
Belletristik	0%	16,67%	2,78%
Rede	0%	0%	0%
WissSchrift	0%	0,33%	15,42%

Tabelle 9: Prozentuales Erscheinen dreier lexikalisch-stilistischer Merkmale in Subkorpora, absteigend angeordnet nach Abdeckung von *kriegen*.

Entsprechend gering ist generell das praktische Gewicht stilistischer Frequenzmaße wie *kriegen*, *bloß* und *lediglich* oder von Interjektionen, Kontraktionen, Stretchwörtern usw., weil viele Texte weder das eine noch das andere enthalten; vgl. Tabelle 9.[13]

[13] Diese Verteilung sagt selbstverständlich nichts über die eigentliche Nähe-Distanz-Klassifizierung aus, sondern dient bestenfalls als Indiz für die generelle Brauchbarkeit der Merkmale.

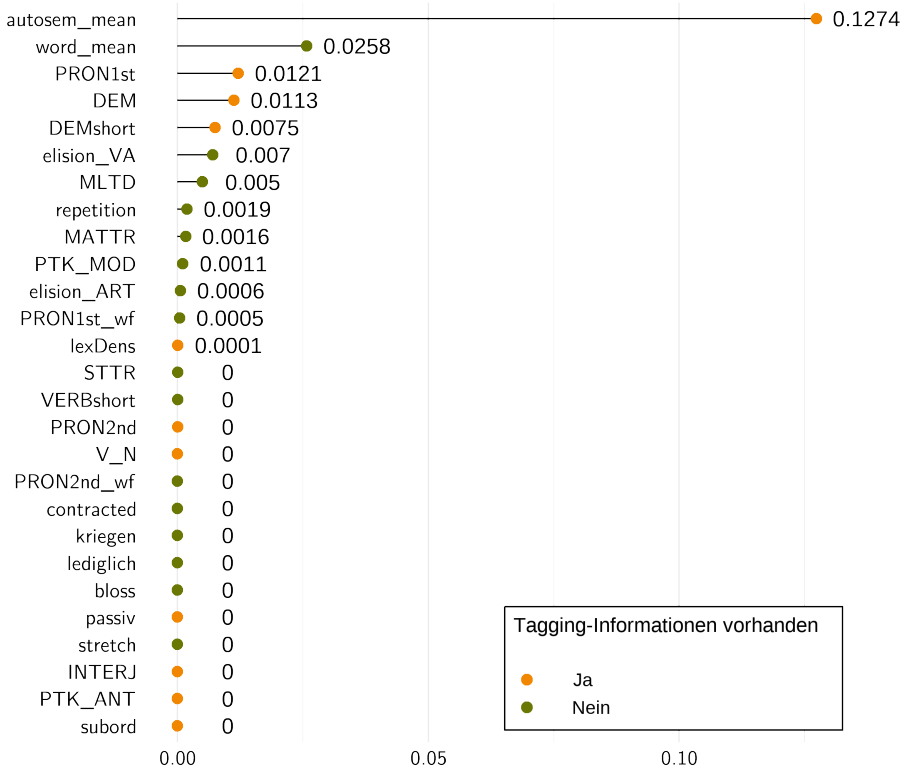


Abbildung 3: Permutation Importance (PI) für Features des Gesamt-Klassifikators.

Auch andere Merkmale können in den Daten diskriminieren, jedoch nicht so zuverlässig und durchgängig wie diejenigen mit hoher PI. Bei K_{Gesamt} spielen beispielsweise die Features `Verbshort` und `PRON2nd_wf` keine Rolle, bei $K_{Oberfläche}$ – der insgesamt weniger Features in Betracht zieht – nehmen sie dagegen nachweisbar Einfluss.

5.3 Nicht-polare Texte im Kontinuum

Nur in vergleichsweise wenigen Fällen unterscheiden sich die Klassifikationen der Modelle markant (vgl. Unterabschnitt 4.5); dies betrifft in erster Linie Songtexte. In anderen Fällen (Interviews, Social Media) hätte man intuitiv vielleicht eine höhere Tendenz zu Nähe/Mündlichkeit erwartet.

Für Interviews analysieren wir nachfolgend anhand der Permutation Importance, wie die mehrheitliche Zuordnung zu Distanz/Schriftlichkeit zustande kommt. Für die unterschiedlich langen Social-Media-Texte gehen wir kurz auf mögliche Gründe variierender Klassifikationen ein. Songtexte betrachten wir zudem aus diachroner und Musikgenre-Perspektive.

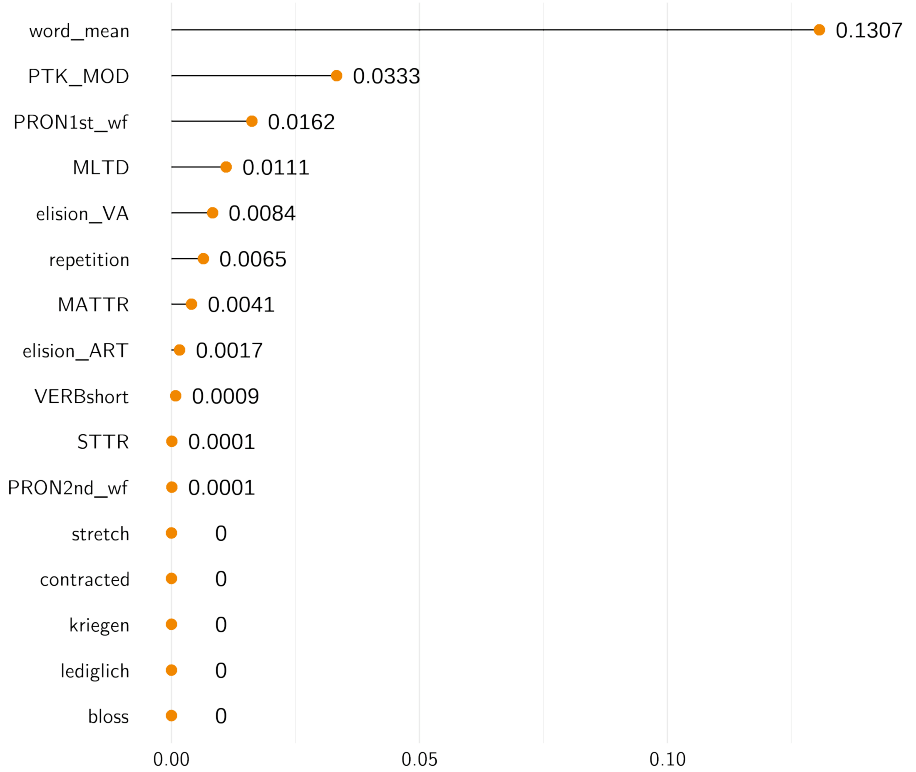


Abbildung 4: Permutation Importance (PI) für Features des Oberflächen-Klassifikators.

5.3.1 Interviews

Interviews im Untersuchungskorpus werden zu ca. 95% als konzeptionell schriftlich klassifiziert, wobei Tokenlänge-Merkmale wie auch bei anderen Textsorten beträchtlichen Einfluss nehmen (Abbildung 3 / Abbildung 4). Ein gezielter Blick auf die Merkmalsverteilung gibt Hinweise auf Hintergründe: Interviews ähneln in ihrer Verteilung der für

das Training eingesetzten Gruppe konzeptionell schriftlicher Texte. Dies verdeutlicht der Violinenplot (Abbildung 5): Je höher die Anzahl der Datenpunkte in einer Spanne, desto breiter ist die „Violine“. Sowohl bei den schriftlichen Distanztexten als auch bei Interviews häufen sich Werte zwischen 7 und 8 (`autosem_mean` (Durchschnitt 7,9/7,32; Median 7,87/7,37)) bzw. um 6 (`word_mean` (Durchschnitt 6,08 (schriftlich)/5,69 (Interview) ; 6,06/5,72)) herum. Bei konzeptionell mündlichen Texten liegen diese Werte eher bei 5 (Durchschnitt 5,15; Median 5,11) bzw. 4 (Durchschnnitt 4,37; Median 4,42).

Darüber hinaus fällt beim gemischten Klassifikator K_{Gesamt} ins Gewicht, dass eher wenige volle Demonstrativa verwendet und Hilfsverben selten elidiert werden. Dies scheint den größeren Anteil gekürzter Demonstrativa und den häufigen Gebrauch von Pronomina der ersten Person Singular zu überwiegen.

Beim Oberflächen-Klassifikator sorgen – neben der Wortlänge und den elidierten Hilfsverben – die seltene Verwendung von Modalpartikeln, eine höhere lexikalische Vielfalt (`MLTD`) sowie seltene unmittelbare Mehrfachnennungen von Wörtern für das tendenziell schriftliche Klassifikationsmuster. Eine plausible (hier nicht weiter beleuchtbare) Ursache liegt mutmaßlich in der für Printpublikationen üblichen Überarbeitung mündlich geführter Gespräche: Ziel solcher Revisionen ist üblicherweise nicht durchweg die exakte Sprecherwiedergabe, sondern eine gewisse mediale Anpassung. Dadurch lassen sich vergleichsweise wenige Wortwiederholungen und Elisionen erklären.

Unter den einflussreichsten Features, die eher in Richtung Nähe/Mündlichkeit steuern, fallen dagegen allein Pronomina der ersten Person ins Gewicht. Vor diesem Hintergrund erschließt sich sich die eindeutige Tendenz beider Klassifikatoren, Interviews mehrheitlich als Texte schriftlicher Konzeption zu klassifizieren.

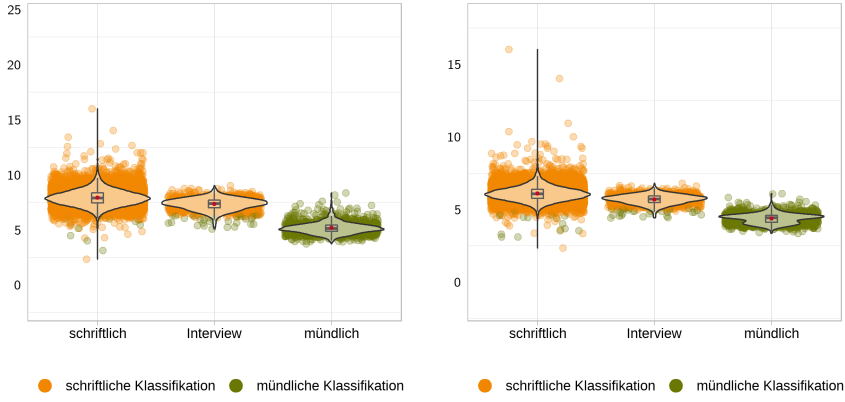
5.3.2 Social Media

Soziale Medien – Tweets und Chats – weisen die mit Abstand kürzesten Textlängen im Untersuchungskorpus auf. Inhalte aus SocialMedia-L mit etwas größeren Textlängen werden von K_{Gesamt} häufiger als konzeptionell mündlich klassifiziert als Inhalte ohne Mindestlänge (siehe Tabelle 8). $K_{Oberfläche}$ macht dagegen insgesamt keinen Unterschied zwischen den beiden Subkorpora.

Der Effekt hängt mutmaßlich damit zusammen, dass es in SocialMedia-L mehr Datenpunkte mit `DEM` > 0 gibt als in den Social-Media-Inhalten ohne Mindesttextlänge, also mehr Tweets und Chats, in denen Demonstrativpronomina überhaupt vorkommen. Da `DEM` für K_{Gesamt} eins der einflussreichsten Merkmale pro Nähe/Mündlichkeit ist (vgl. Abbildung 3), von $K_{Oberfläche}$ jedoch gar nicht beachtet wird, könnte dies die Unterschiede erklären.

Hinsichtlich der bei anderen Textsorten maximal einflussreichen Tokenlänge-Merkmale lassen sich in Abbildung 6 keine markanten Verteilungsunterschiede erkennen. Lediglich die Spannen sind in SocialMedia größer als in SocialMedia-L, Mediane und Mittelwerte unterscheiden sich kaum.

Eine mediale Aufgliederung der Social-Media-Daten zeigt markant unterschiedliche Ergebnisse für Tweets und Chats: Erstere werden unabhängig von Subkorpus oder



(a) `autosem_mean` **(b)** `word_mean`

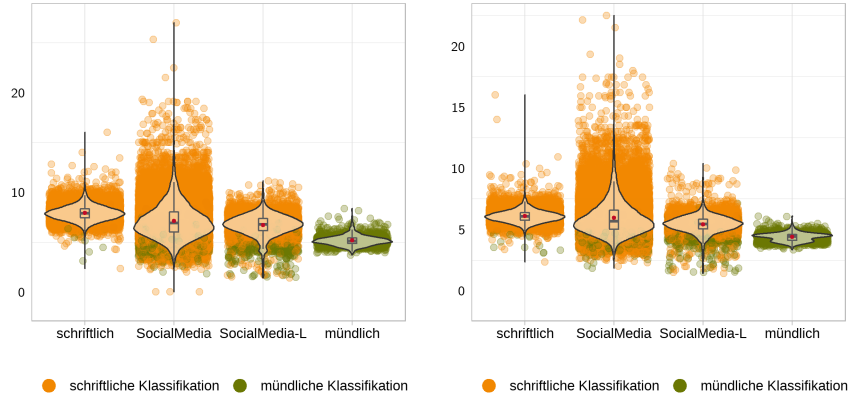
Abbildung 5: Verteilungen der Tokenlängen in Interviews. Es werden alle Datenpunkte (auch Trainingsdaten) abgebildet und entsprechend ihrer Verortung durch den Oberflächen-Klassifikator eingefärbt.

Klassifikationsmodell ganz überwiegend als konzeptionell schriftlich eingestuft. Letztere dagegen zu knapp unter 50% ($K_{Oberfläche}$) bzw. knapp über 50% (K_{Gesamt}) als konzeptionell mündlich. Unter der Prämisse, dass Chats häufiger unter Bekannten ausgetauscht, Tweets dagegen eher an offene Adressatenkreise gerichtet werden, verkörpern Nähe und Distanz die plausibleren Pol-Etiketten als Mündlichkeit und Schriftlichkeit.

5.3.3 Popsongs als „mittige Textsorte“

Songtexte werden von den beiden Klassifikatoren in unterschiedlichem Maß als konzeptionell mündlich bzw. schriftlich eingestuft (Tabelle 8): K_{Gesamt} verteilt beide Konzeptionen ungefähr gleich häufig, $K_{Oberfläche}$ verortet Songs signifikant häufiger als mündliche Nähetexte.

Wirkungsweisen einzelner Merkmale lassen sich auch hier am Beispiel von `DEM` herausstellen: K_{Gesamt}, der das Verhältnis von Demonstrativpronomina zu allen Wörtern eines Texts prominent für dessen Klassifizierung heranzieht, weist Songtexte ohne Demonstrativa (`DEM` $= 0$) verstärkt als konzeptionell schriftlich aus (Abbildung 7, links). Dies korreliert mit der aus den Violinplots ersichtlichen Dichteverteilung der Vergleichsgruppen: Einen Wert von 0 gibt es häufiger bei konzeptionell schriftlichen Texten, während ein höherer Wert häufiger in konzeptionell mündlichen Texten vorkommt. Im Gegensatz dazu, dass der PI-Wert von `DEM` bei K_{Gesamt} relativ hoch ist, wird dieses Merkmal von $K_{Oberfläche}$ nicht berücksichtigt, entsprechend häufiger kommt es dann zur Nähe/Mündlichkeits-Klassifizierung (Abbildung 7, rechts).



(a) `autosem_mean` **(b)** `word_mean`

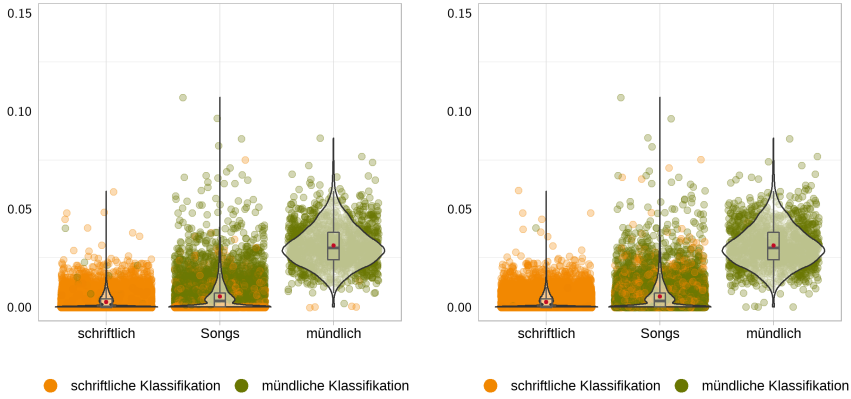
Abbildung 6: Violinplots der Tokenlänge-Merkmale in den Social-Media-Subkorpora. Daten sind auf Basis des Oberflächen-Klassifikators eingefärbt.

Trotz solcher Unterschiede verorten beide Modelle Popsongs (mit einer bewusst weiten Auslegung von „Pop“) unisono als „mittige“ Textsorte. In keinem anderen Subkorpus verteilen sich die betrachteten Merkmale ähnlich ausgewogen; nur ein einziges Merkmal (`kriegen`) findet sich überhaupt nicht im Subkorpus. Tabelle 10 informiert über minimale und maximale Merkmalswerte und gibt jeweils den Mittelwert über alle Songtexte an.

Für eine Plausibilitätsbewertung der Textklassifikationen lohnt sich ein Blick in die Primär- und Metadaten. Dabei fallen Besonderheiten einzelner Künstler bzw. musikalischer Genres (im Songkorpus als Archive recherchierbar; vgl. `https://songkorpus.de`.) ins Auge:

- Klassifikator K_{Gesamt}, der im Songkorpus insgesamt ca. 53% Nähetexte ermittelt, ordnet „traditionelle“ Singer/Songwriter signifikant häufiger unter Distanz/Schriftlichkeit ein.[14] Am deutlichsten trifft das auf Texte von Hannes Wader zu, die zwischen Poesie und erzählender Prosa pendeln (129 konzeptionell schriftlich, 67 konzeptionell mündlich). Ebenfalls als mehrheitlich schriftsprachlich werden die Werke von Reinhard Mey und Tocotronic eingeordnet und interessanterweise mit Herbert Grönemeyer ein Künstler, bei dem erklärtermaßen die Musik vor dem Text kommt, der also beim Komponieren zunächst mit Nonsensphrasen arbeitet und die finalen Texte erst zur fertigen Melodie formuliert.

[14] Die oben angesprochene Tendenz von $K_{Oberfläche}$, Songtexte eher als konzeptionell mündliche Nähetexte zu verorten, fällt hier übrigens schwächer aus, was die Ergebnisse von K_{Gesamt} für das Subsample stützt.



(a) Datenpunkte nach Klassifikationen des Gesamt-Klassifikators eingefärbt.

(b) Datenpunkte nach Verortung des Oberflächen-Klassifikators eingefärbt.

Abbildung 7: Violinplots für das Merkmal `DEM`. Songtexte, bei denen `DEM` gleich oder um 0 ist, werden vom Gesamt-Klassifikator als konzeptionell schriftlich eingestuft. Der Oberflächen-Klassifikator beurteilt sie als konzeptionell mündlich.

- Derselbe Klassifikator gruppiert auch ein Datensubset von 500 Songtexten aus der damaligen DDR, die ein breites Spektrum künstlerischen Schaffens in Ostdeutschland zwischen 1970 und 1990 abdecken, zu zwei Dritteln in die Kategorie Distanz/Schriftlichkeit. Interpretatorisch ließe sich hier vermuten, dass eventuell ein zentral verordneter „künstlerischer Anspruch“ mitspielt, vielleicht auch der Versuch, mit metaphorischen Mitteln Widerstand zu transportieren bzw. Zensurbeschränkungen durch elaborierte sprachliche Kniffe zu umgehen.

- Bemerkenswert erscheint die Einordnung von ebenfalls deutlich über 60% eines 500 Songs umfassenden Datensubsets „Neue Deutsche Welle (NDW)“ von Ende der 1970er bis Mitte 1980er Jahre als schriftsprachliche Distanztexte. Unter dem Gesichtspunkt, dass NDW als Gegenbewegung zu „emotionalen“ Mainstream-Genres entstand und nicht wenige NDW-Bands sprachliche Kühle und Minimalismus als Stilmittel einsetzten, bietet sich diese modellbasierte Einordnung als Ausgangspunkt für anknüpfende Fragestellungen an.

- Schlüssig erscheint aufgrund der seinerzeitigen Popularität von NDW eine insgesamt vermehrte Distanz/Schriftlichkeit in Chartsongs der 1980er Jahre. Abbildung 8 untermauert diese Vermutung. Das zeitlich stratifizierte Subsample enthält sämtliche in den Top-100-Singlecharts platzierten deutschsprachigen Songtexte seit 1970. Es umfasst nicht zwangsläufg gleich viele Inhalte pro Jahr: Da Hitpara-

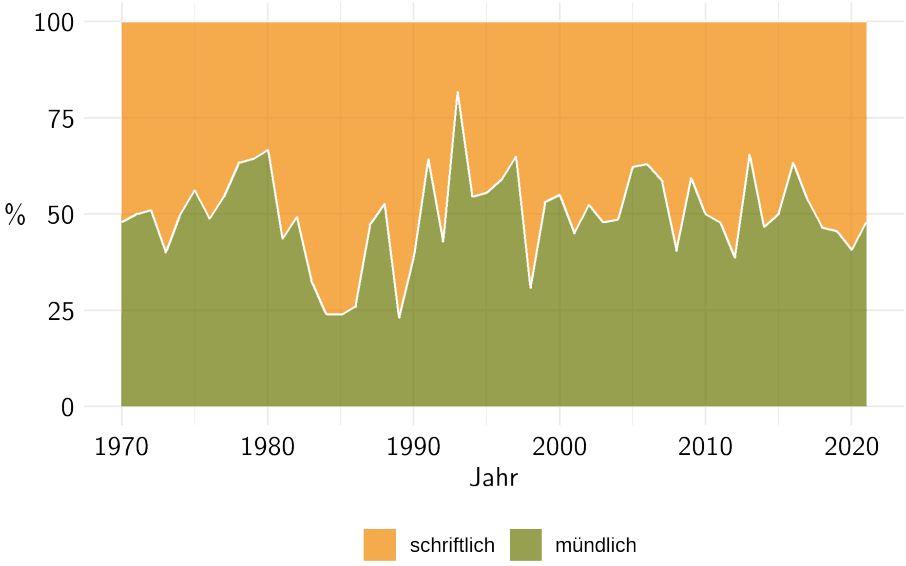


Abbildung 8: Nähe/Mündlichkeit und Distanz/Schriftlichkeit diachron in Chartsongs. Die grüne Fläche stellt den Anteil der als konzeptionell mündlich klassifizierten Texte pro Jahr dar, die orangene den der als konzeptionell schriftlich klassifizierten Texte.

den Moden und Trends reflektieren – also auch eine über die Jahre schwankende Popularität deutschsprachiger Popsongs – ist das Charts-Sample in diesem Sinne zwar repräsentativ, diachron aber nur bedingt ausgewogen.

- Ein umgekehrtes Bild und mit ca. 60% signifikant mehr Nähe als Distanz bietet zeitgenössischer Hiphop-Sprechgesang. Hierzu inspizieren wir ein Subsample mit 1000 Deutschrap-Songs (564 zu 436) sowie speziell Texte der Band 'Fettes Brot' (82 zu 53). Die Tendenz erscheint plausibel: Hiphop spricht in der Hauptsache ein junges Publikum an und gilt mit seinen Freestyle-Elementen, parataktischen Formen, Ellipsen usw. als vergleichsweise nahe an Umgangssprache. Interessanterweise ist Hiphop das einzige Musikgenre im Korpus, bei dem der Verzicht auf Tagging-Infos zu (marginal) weniger Nähe/Mündlichkeitsklassifikationen führt: 540 zu 459 ($K_{Oberfläche}$) vs. 563 zu 436 (K_{Gesamt}).

	Minimum	Maximum	Mittelwert
STTR	0	0,831	0,303
MATTR	0,025	0,970	0,328
MLTD	4,836	1357,235	115,693
PRON1st_wf	0	0,316	0,050
PRON2nd_wf	0	0,209	0,027
bloss	0	0	0
lediglich	0	0,004	0,000008
kriegen	0	0	0
word_mean	3,088	9,208	4,559
word_med	2	8	4,015
elision_ART	0	1	0,239
elision_VA	0	1	0,176
VERBshort	0	1	0,543
contracted	0	0,133	0,004
PTK_MOD	0	0,234	0,034
repetition	0	0,371	0,002
stretch	0	0,055	0,0002
lexDens	0,204	0,678	0,450
DEM	0	0,107	0,005
DEMshort	0	0,107	0,004
PRON1st	0	0,193	0,043
PRON2nd	0	0,202	0,023
V_N	0	51	0,799
autosem_mean	3,264	10,339	5,541
autosem_med	2	8,5	5,063
INTERJ	0	0,368	0,003
PTK_ANT	0	0,281	0,001
passiv	0	0,280	0,001
subord	0	2,182	0,140

Tabelle 10: Feature-Werte in Songtexten.

6 Zusammenfassung und Ausblick

Prototypische, konzeptionell mündliche Nähetexte lassen sich anhand unseres – für diese erste Annäherung notwendigerweise opportunistisch zusammengestellten – Featuresets sehr gut von konzeptionell schriftlichen Distanztexten unterscheiden. Wir haben hierfür Phänomene aufgegriffen, die in der sprachwissenschaftlichen Forschung als Markierer für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit diskutiert werden, und die ohne tiefe syntaktische Analyse („deep parsing“) maschinell bestimmbar sind. Aufbauend auf der sehr guten binären Klassifikation haben wir die trainierten statistischen Modelle auf weitere Textsorten angewendet. Daraus ergibt sich keine unmittelbar linguistisch plausible Positionierung dieser Textsorten im Kontinuum. Allerdings lassen sich wertvolle Aussagen darüber ableiten, wie sich die herangezogenen Merkmale in situativ und medial heterogenen Sprachdaten verteilen und von den Klassifikatoren genutzt werden. Unser Verfahren ersetzt keinen systematischen theoretischen Überbau und keine ausdifferenzierte Methodik zur Beurteilung von Nähe-Distanz bzw. Mündlichkeit-Schriftlichkeit. Aber es illustriert die Praktikabilität empirischer Verfahren als evidenzbasierte Hilfestellung bei der Evaluation; die Auswahl bekannter Marker aus der linguistischen Forschung ermöglicht eine Einbettung statistischer Erkenntnisse in die Sprachtheorie.

Interessant erscheint die maschinelle Verortung von Popsongs als „mittige Textsorte“, die viele Nähe-Distanz-Merkmale vereint und den Gegenstandsbereich damit als hochattraktiv für die empirisch-deskriptive Sprachforschung ausweist.

Unsere breit stratifizierte Textsorten-Analyse untermauert bekannte Argumente, warum sich Nähe und Distanz als Pole zur Beschreibung von Äußerungsformen besser eignen als der mediale Dualismus Mündlichkeit/Schriftlichkeit: Wir konnten empirisch belegen, dass unsere Modelle mündlich kommunizierte Äußerungen wie Interviews oder Reden summa summarum vollkommen anders beurteilen als Telefon- oder Unterrichtsgespräche, die unter verschiedengestaltigen Kommunikationsbedingungen ablaufen und entsprechend divergierende Versprachlichungsstrategien wählen. Die Plausibilität des Nähe-Distanz-Konzepts wird unterstrichen durch die – nach Aufgliederung der Social-Media-Daten – aufgedeckten Klassifikationsunterschiede für Tweets und Chats.

Offenkundig besitzen Wortlänge-Features für sich genommen in den trainierten Modellen bereits eine hervorragende diskriminierende Vorhersagekraft; hier könnten Folgeuntersuchungen das Gewicht der übrigen Merkmale bei Wegfall dieser prominenten Markierer präziser ausloten. Dabei gälte es, weitere potenziell redundante Maße hinsichtlich ihrer wechselseitigen Abhängigkeit zu analysieren, mit dem Ziel, beispielsweise nur noch ein Maß für Aussagen zur lexikalischen Vielfalt (bislang: MLTD, MATTR, STTR) heranzuziehen. Einzelne Features bieten offenkundiges Optimierungspotenzial, etwa die Messung von Passivstil oder von Wortwiederholungen, bei denen außer Einzelwörtern auch Wortfolgen eine Rolle spielten sollten.

Ein naheliegendes Desiderat besteht in der Anreicherung unseres Textsortenspektrums um weitere bislang nicht berücksichtigte Textsorten. Auch die methodische Einbeziehung von Streuungs-/Dispersionsmaßen zur Bewertung der Verteiltheit von Merkmalen steht noch aus. Zur Überprüfung der Bewertungsgüte bzw. des Feature-Status bleibt ein breit

stratifizierter annotierter Goldstandard wünschenswert, in den Urteile linguistischer Experten einfließen. Damit verbindet sich die spannende Frage, wie sich eher qualitative Ansätze und Theorien mit maschinengestützten Verfahren abgleichen lassen. Vor diesem Hintergrund stellen wir unsere trainierten Klassifikationsmodelle wissenschaftsöffentlich zur Reproduktion, Evaluierung und Optimierung zur Verfügung.[15]

Literatur

Ágel, V. & Hennig, M. (2006a). Praxis des Nähe- und Distanzsprechens. In V. Ágel & M. Hennig (Hrsg.), *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000* (S. 33-74). Tübingen: Niemeyer.

Ágel, V. & Hennig, M. (2006b). Theorie des Nähe- und Distanzsprechens. In V. Ágel & M. Hennig (Hrsg.), *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000* (S. 3-31). Tübingen: Niemeyer.

Androutsopoulos, J. K. (2003). Online-Gemeinschaften und Sprachvariation : soziolinguistische Perspektiven auf Sprache im Internet. *Zeitschrift für germanistische Linguistik : deutsche Sprache in Gegenwart und Geschichte*, *31* (2), 173 – 197.

Barbour, S. & Stevenson, P. (1998). *Variation im Deutschen. Soziolinguistische Perspektiven.* Berlin, Boston: De Gruyter.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, *41* (1), 161–164. Zugriff auf `https://doi.org/10.1515/zgl-2013-0009`

Biber, D. & Conrad, S. (2019). *Register, genre, and style* (2. Aufl.). Cambridge University Press. doi: 10.1017/9781108686136

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The longman grammar of spoken and written english.* Harlow, UK.: Longman.

Cotgrove, L. A. (2017). *Is Computer-Mediated Communication "written Colloquial Speech" (Kilian 2001)? A Quantitative Study of German-Language YouTube Comments.* Nottingham: University of Nottingham. (MA Thesis)

Covington, M. & McFall, J. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*, 94-100. Zugriff auf `https://doi.org/10.1080/09296171003643098`

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A. & Swinton, J. (2019). xtable: Export Tables to LaTeX or HTML [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=xtable`

Debeer, D., Hothorn, T. & Strobl, C. (2021). permimp: Conditional Permutation Importance [Software-Handbuch].

Dürscheid, C. (2016). *Einführung in die Schriftlinguistik. Mit einem Kapitel zur Typographie von Jürgen Spitzmüller* (5. Aufl.). Göttingen: Vandenhoeck & Ruprecht.

Eichinger, L. M. (2017). Gesprochene Alltagssprache. In Deutsche Akademie für Sprache und Dichtung/Union der deutschen Akademien der Wissenschaften (Hrsg.),

[15] `https://songkorpus.de/data/`

Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache (S. 283 – 331). Tübingen: Stauffenburg.

Feilke, H. & Hennig, M. (Hrsg.). (2016). *Zur Karriere von ›Nähe und Distanz‹: Rezeption und Diskussion des Koch-Oesterreicher-Modells.* Berlin: De Gruyter.

Grammis. (2018). *Abtönungspartikeln.* Mannheim: Leibniz-Institut für Deutsche Sprache. Zugriff auf `https://grammis.ids-mannheim.de/systematische-grammatik/1322` (Grammatisches Informationssystem: Systematische Grammatik)

Grammis. (2020). *Elision.* Mannheim: Leibniz-Institut für Deutsche Sprache. Zugriff auf `https://grammis.ids-mannheim.de/terminologie/1169` (Grammatisches Informationssystem: Wissenschaftliche Terminologie)

Halliday, M. (1985). *Spoken and Written Language.* Geelong: Deakin University Press.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. & van der Laan, M. (2006). Survival Ensembles. *Biostatistics*, *7* (3), 355–373.

Kilian, J. (2001). T@stentöne. Geschriebene Umgangssprache in computervermittelter Kommunikation. In M. Beißwenger (Hrsg.), *Chatkommunikation.* (S. 55–78). Stuttgart: Ibidem.

Kleiner, S., Berend, N., Knöbl, R. & Brinckmann, C. (2014). „Deutsch heute“: ein sprachgebietsweites Forschungsprojekt zur regionalen Variation in der gesprochenen deutschen Standardsprache. *Klagenfurter Beiträge zur Sprachwissenschaft*, *34-36*, 179 – 193. Zugriff auf `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-28744`

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36* (1), 15–43. Zugriff auf `https://doi.org/10.1515/9783110244922.15`

Koch, P. & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, *35*, 346 - 375.

Kuhn, M. (2022). caret: Classification and regression training [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=caret` (R package version 6.0-93)

Kupietz, M., Lüngen, H., Kamocki, P. & Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In N. Calzolari et al. (Hrsg.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Miyazaki, Japan: ELRA.

Margaretha, E. & Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics (JLCL)*, *29* (2), 59–82. Zugriff auf `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-33306`

Mccarthy, P. & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*, 381-92. Zugriff auf `https://doi.org/10.3758/BRM.42.2.381`

Meier-Vieracker, S. (2018). Fussball digital - korpuslinguistische Perspektiven auf die Sprache des Fußballs. *Sprachreport*, *34* (2), 1-9. Zugriff auf `https://nbn`

-resolving.org/urn:nbn:de:bsz:mh39-75266

Ortmann, K. & Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. In *Proceedings of the sixth workshop on NLP for similar languages, varieties and dialects* (S. 64-79). Ann Arbor, Michigan: Association for Computational Linguistics. Zugriff auf https://aclanthology.org/W19-1407

Ortmann, K. & Dipper, S. (2020). Automatic orality identification in historical texts. In *Proceedings of the 12th language resources and evaluation conference* (S. 1293-1302). Marseille, France: European Language Resources Association. Zugriff auf https://aclanthology.org/2020.lrec-1.162

R Core Team. (2022). R: A Language and Environment for Statistical Computing [Software-Handbuch]. Zugriff auf https://www.R-project.org/

Rehm, G. (2002). Schriftliche Mündlichkeit in der Sprache des World Wide Web. In A. Ziegler & C. Dürscheid (Hrsg.), *Kommunikationsform E-Mail* (Bd. 7, S. 263–308). Tübingen: Stauffenburg.

Schirrmeister, L., Rummel, M., Heine, A., Suppus, N. & Mendoza Sánchez, B. (2021). Gingko – ein Korpus der ingenieurwissenschaftlichen Sprache. *Deutsch als Fremdsprache* (4). Zugriff auf https://doi.org/10.37307/j.2198-2430.2021.04.04

Schlobinski, P. (2005). Mündlichkeit/Schriftlichkeit in den Neuen Medien. In L. M. Eichinger (Hrsg.), *Standardvariation. Wie viel Variation verträgt die deutsche Sprache?* (S. 126 – 142). Berlin: de Gruyter.

Schmidt, T. (2017). DGD – die Datenbank für Gesprochenes Deutsch. *Zeitschrift für germanistische Linguistik*, *45* (3), 451-463. Zugriff auf https://doi.org/10.1515/zgl-2017-0027

Schmidt, T. (2018). Gesprächskorpora. In M. Kupietz & T. Schmidt (Hrsg.), *Korpuslinguistik* (S. 209-230). Berlin: De Gruyter.

Schneider, R. (2019a). *Corpus of Song Lyrics.* Mannheim. Zugriff am 19.03.2023 auf https://songkorpus.de

Schneider, R. (2019b). "Konservenglück in Tiefkühl-Town"– Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)* (S. 229-236). Erlangen: German Society for Computational Linguistics (GSCL). Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93189

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus". In N. Calzolari et al. (Hrsg.), *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)* (S. 842-848). Paris: European Language Resources Association. Zugriff auf http://nbn-resolving.de/urn:nbn:de:bsz:mh39-100347

Schneider, R. (2022). Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. *Sprachreport*, *38* (1), 38-50. Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-109499

Storrer, A. (2000). Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet. In G. G. Voš, W. Holly & K. Boehnke (Hrsg.), *Neue Medien im Alltag. Begriffsbestimmungen eines interdisziplinären*

Forschungsfeldes (S. 151-175). Opladen: Leske + Budrich.

Strobl, C., Boulesteix, A.-L., Kneib, T., Thomas, A. & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics published by Springer Nature*, *9* (307), 1–11. Zugriff auf https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, *8* (25). Zugriff auf https://doi.org/10.1186/1471-2105-8-25

Werner, V. (2021). Catchy and conversational? : a register analysis of pop lyrics. *Corpora : corpus-based language learning, language processing and linguistics*, *16* (2), 237–270. Zugriff auf https://fis.uni-bamberg.de/handle/uniba/51410

Westpfahl, S., Schmidt, T., Jonietz, J. & Borlinghaus, A. (2017). *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Version 1.1* (Working Paper). Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-60634

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis [Software-Handbuch]. New York: Springer-Verlag. Zugriff auf https://ggplot2.tidyverse.org

Zifonun, G., Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache.* Berlin/New York: De Gruyter. Zugriff auf https://doi.org/10.1515/9783110872163

Korrespondenzanschrift

Sarah Broll
Leibniz-Institut für Deutsche Sprache
broll@ids-mannheim.de

Roman Schneider
Leibniz-Institut für Deutsche Sprache
schneider@ids-mannheim.de