

Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Medieninformatik

Diskursmarker in schriftlichem und akustischem Diskurs

Bachelorarbeit

Johanna Sacher
geb. am: 09.04.1994 in Herford

Matrikelnummer 117353

1. Gutachter: Prof. Dr. Benno Stein
2. Gutachter: PD Dr. Andreas Jakoby

Datum der Abgabe: 14. Januar 2021

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, 14. Januar 2021

.....
Johanna Sacher
.....

Zusammenfassung

Diese Bachelorarbeit vergleicht die Nutzung von Diskursmarkern in schriftlichem und akustischem Diskurs. Diskursmarker sind Wörter wie *and*, *so* und *but*, die an sich keine Bedeutung zum Satz beitragen, aber die Beziehung zwischen zwei Diskurssegmenten signalisieren und als Wegweiser im Text dienen. Die grundlegende Hypothese dieser Arbeit ist, dass Diskursmarker in schriftlichem und akustischem Diskurs unterschiedlich verwendet werden, mit entsprechenden Konsequenzen beispielsweise für Schreibassistenz-Programme sowie Programme zur Text- und Sprachgenerierung. Um diese Hypothese zu untersuchen, wurde das Vorkommen von Diskursmarkern in vier verschiedenen Corpora analysiert. Zwei dieser Corpora enthalten ursprünglich schriftliches Material, die anderen beiden Transkripte von Audiomaterial. Anhand dieser Datengrundlage wird die Verwendung von Diskursmarkern in schriftlichem und akustischem Diskurs sowie innerhalb verschiedener Genres und Konversationsarten untersucht und verglichen. Zu diesem Zweck wurde ein Programm entwickelt, das Diskursmarker in einer großen Menge unstrukturierter Texte erkennen und annotieren kann. Die resultierenden Verteilungswerte wurden visualisiert und statistisch ausgewertet. Die Ergebnisse zeigen, dass Diskursmarker in akustischem Diskurs vermehrt eingesetzt werden und vor allem auch in interaktiven Diskursformen eine große Rolle spielen.

Inhaltsverzeichnis

| | |
|--|-----------|
| Glossar | 1 |
| Abkürzungen | 2 |
| 1 Einleitung | 3 |
| 2 Grundlagen & Verwandte Arbeit | 7 |
| 2.1 Oraler und Literater Diskurs | 7 |
| 2.2 Readability und Listenability | 8 |
| 2.3 Ratgeberliteratur zu Listenability | 10 |
| 3 Diskursmarker | 12 |
| 3.1 Begriff | 12 |
| 3.2 Einteilung in Bedeutungsgruppen | 15 |
| 4 Einteilung der Textdaten in Textsorten | 22 |
| 4.1 Diskursarten | 22 |
| 4.2 Genres | 23 |
| 4.3 Konversationsarten | 23 |
| 5 Corpora | 25 |
| 5.1 Akustische Corpora | 25 |
| 5.2 Schriftliche Corpora | 39 |
| 6 Feature Engineering | 42 |
| 6.1 Aufbau der Analysis Pipeline | 42 |
| 6.2 Diskursmarker erkennen durch String-Matching | 44 |

| | |
|---|------------|
| 6.3 Scores | 46 |
| 7 Auswertung | 51 |
| 7.1 Welche Textsorten stützen sich besonders auf Diskursmarker? | 52 |
| 7.2 An welchen Positionen im Text stützen sich die jeweiligen Textsorten besonders auf Diskursmarker? | 58 |
| 7.3 An welchen Positionen im Satz stützen sich die jeweiligen Textsorten besonders auf Diskursmarker? | 61 |
| 7.4 Auf welche Klassen von Diskursmarkern stützen sich die jeweiligen Textsorten besonders? | 64 |
| 7.5 Welche Diskursmarker werden innerhalb der jeweiligen Klassen besonders genutzt? | 68 |
| 8 Schlussfolgerung | 74 |
| 9 Ausblick | 77 |
| A Tabellen | 78 |
| B Diagramme und Statistiken | 89 |
| Literaturverzeichnis | 106 |

Danksagung

Vielen Dank an Prof. Dr. Matthias Hagen und Johannes Kiesel für die Bereitstellung dieses Themas, auch wenn es sich im Laufe der Zeit ein wenig gewandelt hat. Mein Dank gilt außerdem Dr. Khalid Al-Khatib und im Besonderen Johannes Kiesel für die anhaltende Unterstützung während dieser Arbeit.

Vielen Dank außerdem an Prof. Dr. Benno Stein und PD Dr. Andreas Jakoby für das Bewerten meiner Arbeit.

Zudem möchte ich danken:

Meiner Schwester – für das Heraussuchen vergangener Statistik-Vorlesungen.

Christian Paffrath – für Tee und Brote und den Boden unter meinen Füßen.

Valerie Lemuth – für gemeinsames Denken.

Drago – für notwendige Pausen.

Und vor allem meinen Eltern – für alles.

Glossar

Corpus in der Linguistik eine Sammlung von schriftlichem oder gesprochenem Material, mit dessen Hilfe Sprache untersucht werden kann¹.

Diskurs in der Linguistik eine Einheit von Sprache länger als ein einzelner Satz².

In der Diskursanalyse wird gesprochene und geschriebene Sprache untersucht.

Diskursmarker Wörter, die eine Beziehung zwischen Diskurssegmenten anzeigen, unter anderem Bindewörter wie *and*, *but* und *so*, näher erläutert in Kapitel 3.

Homograph Wörter mit gleicher Schreibweise aber unterschiedlicher Bedeutung.

Listenability Maß für die Verständlichkeit akustischer Texte.

Readability Maß für die Verständlichkeit schriftlicher Texte.

UIMA *Unstructured Information Management Applications*; Programme, die eine große Menge an unstrukturierten Informationen verarbeiten und analysieren können, um relevante Informationen herauszuarbeiten³.

Word Error Rate Maß für die Menge an Wortfehlern in der Transkribierung eines akustischen Textes.

¹<https://dictionary.cambridge.org/dictionary/english/corpus>, accessed 9.1.2021

²<https://www.thoughtco.com/discourse-language-term-1690464>, accessed 26.12.2020

³<https://uima.apache.org/>, accessed 1.1.2021

Abkürzungen

CSV Comma-separated values, Datenformat.

DM Diskursmarker, s. Kapitel 3.

EG Effektgröße, s. Kapitel 7.

JSON JavaScript Object Notation, Datenformat.

LS literat-schriftlich, s. Abschnitt 2.1.

OA oral-akustisch, s. Abschnitt 2.1.

STM Exchange-Streaming-Media-Datei, Datenformat.

TXT Text File, Datenformat.

WER Word Error Rate, s. Unterunterabschnitt 5.1.2.2.

XMI XML Metadata Interchange, Datenformat.

XML Extensible Markup Language, Auszeichnungssprache zur Codierung von Informationen.

1

Einleitung

Der Redakteur schreibt immer für die Ohren, er muss sich immer als Sprecher denken. Wir »hören« immer.

Schneider and Rau [2012]

Sprachassistenten wie Alexa¹ und Siri² gehören immer häufiger zum alltäglichen Leben. Sie führen verschiedenste Aufgaben aus, beantworten Fragen und halten die nutzende Person über das Weltgeschehen auf dem Laufenden. Und in der Theorie mag es praktisch klingen, sich beispielsweise den Leitartikel der aktuellen Tageszeitung vorlesen zu lassen, während man das Frühstück vorbereitet. Doch wurde dieser Artikel – in den meisten Fällen – geschrieben, um gelesen zu werden. Wird er nun vorgelesen und angehört, kann es sein, dass er auf diesem Wege nicht mehr so leicht verständlich ist. Zuhörer könnten die Nachrichten missverstehen oder gar frustriert ausschalten, weil das Gehörte zu kompliziert und unverständlich scheint.

Klassische Zeitungsartikel werden nicht mit dem primären Ziel verfasst, gehört zu werden. Ihr ursprüngliches Medium ist die Schrift, nicht die Akustik. Werden sie vorgelesen, bleibt das Konzept zwar dasselbe, jedoch ändert sich das Medium: Das literate Konzept des für das schriftliche Medium verfassten Zeitungsartikels trifft hier auf das akustische Medium des vorlesenden Sprachassistenten. Was fehlt ist die orale Konzeption des Textes für dieses Medium. Aufgrund der Unterschiede zwischen schriftlichem und akustischem Diskurs

¹<https://developer.amazon.com/en-US/alexa>

²<https://www.apple.com/siri/>

muss auch die Verständlichkeit des schriftlichen Diskurses, hier genannt *Readability*, unterschieden werden von der Verständlichkeit des akustischen Diskurses, hier genannt *Listenability*. Schriftlicher Diskurs setzt keine bestimmte Lesegeschwindigkeit voraus, was die Verständlichkeit auch komplexerer Strukturen ermöglicht. So können beispielsweise Teile des Textes erneut gelesen werden, um Sachverhalte besser nachzuvollziehen. Akustischer Diskurs hingegen wird in Echtzeit verarbeitet und die Verständlichkeit hängt daher stark von der Verarbeitungskapazität der hörenden Person ab [Ortmann and Dipper, 2019].

Da aber das Ziel von Kommunikation immer Verständigung sein sollte und Verständigung nur aus Verständnis folgen kann, muss, wenn eine parallele Nutzung von ursprünglich zum Lesen gedachter Texte über das akustische und das schriftliche Medium ermöglicht werden soll, ein Mindestmaß an Listenability dieser Texte gewährleistet werden. Eine Einschränkung der Readability ist dabei nicht zu befürchten, denn »Die Schrift ist eine spätere Zugabe zur mündlichen Rede und das stumme Lesen wiederum ein sehr spätes Stadium ihrer Nutzung.« [Schneider and Raue, 2012]. Diese »orale Tradition« [Horne, 2019a] des Schreibens und Lesens hat zur Folge, dass Listenability stets Readability impliziert; »Laut lesen ist für den Autor die beste Qualitätskontrolle.« [Schneider and Raue, 2012]. Um die Listenability literater Texte gewährleisten zu können, muss untersucht werden, welche Eigenschaften ein Text aufweist, der speziell für das akustische Medium verfasst wurde. Von der Fachliteratur werden hier unter anderem einfache Worte und Sätze [Schneider and Raue, 2012], bildhafte Beschreibungen, sowie der gezielte Einsatz von *Diskursmarkern* empfohlen [Horne, 2019a, McCormick, 2013].

Die Empfehlung zu kurzen Sätzen und einfachen Worten ist relativ leicht zu realisieren: lange Sätze werden aufgeteilt in kürzere, und komplizierte Worte durch einfache Synonyme ersetzt. Beides lässt sich messen und überprüfen, beispielsweise an der Satzlänge oder der generellen Häufigkeit eines Wortes. Der bewusste Einsatz von Diskursmarkern hingegen ist nicht durch die einfache Anwendung von Regeln möglich. Diskursmarker, unter anderem Bindewörter wie *and*, *but* und *so* (siehe Kapitel 3), beeinflussen die Struktur des Textes, sie sind der Klebstoff oder der Abstandshalter; der Wegweiser im Text. Sie zu verwenden ist McCormick [2013] zufolge eine Kunst: »One aspect of the craft of writing in-

volves learning how to use connective words or phrases, and also knowing when to avoid them.». Diese Schwierigkeit macht sie zu einem interessanten Aspekt der Listenability-Forschung. Wie werden Diskursmarker in schriftlichen Texten verwendet und wie in Texten für das akustische Medium? Lassen sich Unterschiede feststellen, die genutzt werden können, um die Listenability von Texten zu gewährleisten? Gibt es beispielsweise in akustischen Texten mehr Diskursmarker als in schriftlichen und könnte ihr Einsatz daher aufgrund von Hörgewohnheiten einen Text für das akustische Medium besser verständlich gestalten?

Im Rahmen dieser Arbeit soll empirisch untersucht werden, wo welche Diskursmarker in verschiedenen Diskursarten auftreten. Dazu wird erstens unterschieden zwischen literat-schriftlichem und oral-akustischem Diskurs, zweitens zwischen verschiedenen Genres des oral-akustischen Diskurses und drittens zwischen Konversationsarten des oral-akustischen Diskurses. Auf Grundlage dieser Betrachtungen sollen Rückschlüsse darauf gezogen werden, welche Rolle Diskursmarker für die Listenability eines Textes spielen.

Anhand der Fragestellungen

1. Welche Textsorten stützen sich besonders auf Diskursmarker?
2. An welchen Positionen im Text stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?
3. An welchen Positionen im Satz stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?
4. Auf welche Klassen von Diskursmarkern stützen sich die jeweiligen Textsorten besonders?
5. Welche Diskursmarker werden innerhalb der jeweiligen Klassen besonders genutzt?

bringt diese Arbeit, gestützt auf eine empirische Auswertung der Diskursmarkernutzung in vier Datensätzen, mehrere neue Erkenntnisse hervor. Insbesondere wird Evidenz dafür geliefert, dass (A) Diskursmarker in oral-akustischem Diskurs häufiger eingesetzt werden als in literat-schriftlichem; (B) Interaktiveren Diskursformen Diskursmarker eher am Anfang einzusetzen, während sachlichere Texte eher am Ende auf Diskursmarker zurückgreifen; (C) Diskursmarker wie *and*, die das Hinzufügen eines weiteren Aspekts signalisieren, die am häufigs-

ten genutzten Diskursmarker sind; (D) oral-akustischer Diskurs eher umgangssprachlichere Diskursmarker nutzt als literat-schriftlicher Diskurs, die am meisten genutzten Diskursmarker sich jedoch kaum unterscheiden.

Wichtige theoretische Grundlagen für diese Arbeit sowie verwandte Forschung werden in Kapitel 2 und 3 betrachtet. Kapitel 2 beschäftigt sich dabei mit grundsätzlich bedeutsamen Konzepten, Begriffen und Forschungsarbeiten, während Kapitel 3 genauer auf Diskursmarker eingeht. Diskutiert werden der Begriff *Diskursmarker*, die Funktion von Diskursmarkern in der englischen Sprache sowie eine mögliche Unterteilung in Bedeutungsgruppen. Anschließend werden in Kapitel 4 die Textsorten definiert, die in der Analyse verglichen werden. Kapitel 5 beschreibt, welche Daten für die Analyse genutzt wurden. Es wird auf Probleme bei der Datenfindung und Aufbereitung eingegangen, sowie eine Beschreibung der letztendlich verwendeten Daten gegeben. Um Diskursmarker in einer großen Datenmenge zu finden und ihre Verteilung statistisch auszuwerten, wurde ein Programm geschrieben, das in Kapitel 6 näher erläutert wird. Die aus dieser Analyse resultierenden Ergebnisse werden in Kapitel 7 beschrieben und ausgewertet. In Kapitel 8 werden die Ergebnisse noch einmal kurz zusammengefasst, um einen Überblick zu geben und einen Zusammenhang zu Listenability herzustellen. Abschließend wird in Kapitel 9 ein Blick auf noch offene Fragen und Möglichkeiten für weitere Forschungsansätze geworfen.

2

Grundlagen & Verwandte Arbeit

Im Folgenden werden wichtige theoretische Grundlagen und Konzepte für diese Arbeit sowie verwandte Forschung betrachtet. Abschnitt 2.1 beschreibt, wie Diskurs konzipiert und kommuniziert werden kann und Abschnitt 2.2 geht auf die Verständlichkeit verschiedener Diskursarten ein. Abschnitt 2.2 und Abschnitt 2.3 beschäftigen sich mit den Faktoren, die Verständlichkeit von Diskurs beeinflussen können.

2.1 Oraler und Literater Diskurs

Oral und *Literat* sind Konzepte, die beschreiben, für welches Medium der Diskurs ursprünglich konzipiert wurde. So sind beispielsweise die Kommentare unter einem YouTube¹ Video als oral konzipierter Diskurs einzuordnen, sie wurden jedoch über das schriftliche Medium kommuniziert. Der vorgelesene Zeitungsartikel hingegen wurde für das schriftliche Medium konzipiert, jedoch über das akustische kommuniziert [Ortmann and Dipper, 2019]. Tabelle 2.1 zeigt, wie Konzepte und Medien in diesem Zusammenhang beliebig kombiniert werden können.

Erst kürzlich gab es erste Forschungsansätze zu diesem Thema in der Computerlinguistik. Ortmann and Dipper [2019] stellen in ihrer Arbeit Eigenschaften vor, die *literaten/schriftlichen* von *oralem/akustischem*² Diskurs unterscheiden.

¹<https://www.youtube.com/>

²Ortmann and Dipper nennen das Medium *spoken* (gesprochen), doch wird es in dieser Arbeit aufgrund des Bezugs auf Sprachassistenten *akustisch* genannt.

| Medien | Konzepte | |
|-------------|---|---------------------------------------|
| | Literat | Oral |
| Schriftlich | Stummes Lesen eines Zeitungsartikels | Stummes Lesen von YouTube-Kommentaren |
| Akustisch | Hören eines vorgelesenen Zeitungsartikels | Persönliches Gespräch |

Tabelle 2.1: Zusammenhang von Diskursmedien und Diskurskonzepten an Beispielen

Untersucht wurden die Unterschiede in Referenz und Deixis³, Komplexität, Syntax, grammatischen und lexikalischen Variationen, grafischen Features (beispielsweise die Auslassung einzelner Buchstaben oder wiederholte Ausrufezeichen), und einzelnen Wortgruppen wie Interjektionen oder Partikeln. Vor allem einfache Komplexitätsmaße wie durchschnittliche Satz- und Wortlängen erwiesen sich hier als Schlüsselindikatoren. Es wurde festgestellt, dass gesprochene Sprache weniger komplex ist als geschriebene; die Sätze sind kürzer und weniger verschachtelt, sie beinhalten mehr Wiederholungen und mehr Koordination.

2.2 Readability und Listenability

Mit *Readability* wird im Folgenden das Leseverständnis beschrieben, wie gut also ein Text verständlich ist, wenn er gelesen wird. Die Forschung zu Readability ist extensiv; Dubay [2007] gibt einen guten Überblick über Readability-Literatur im Zusammenhang mit Listenability. Im Laufe dieser Forschungsarbeiten wurden bereits viele Formeln und Skalen entwickelt, um die Readability eines Textes zu messen. Der von Flesch [1948] vorgestellte *Flesch-Reading-Ease*-Index stützt sich auf die durchschnittliche Satzlänge sowie die durchschnittliche Anzahl der Silben pro Wort. Dieser Index ist wegen seiner Einfachheit weit verbreitet und wird beispielsweise auch im Textverarbeitungsprogramm Microsoft Word⁴ be-

³Bezugnahme auf Personen, Orte, Gegenstände oder Zeit durch Worte wie *du*, *dort*, *jenes*, *heute*: <https://de.wikipedia.org/wiki/Deixis>, accessed 29.12.2020

⁴<https://www.microsoft.com/en/microsoft-365/word>, accessed 14.12.2020

nutzt, um die Readability eines Textes zu prüfen⁵. Die Dale-Chall Formel nutzt ebenfalls die durchschnittliche Satzlänge, nimmt aber anstelle der Silbenanzahl die Anzahl der schwierigen Worte zu Hilfe, um daraus ein Schwierigkeitslevel zu berechnen [Chall and Dial, 1948]. In der Listenability-Forschung können diese Formeln also zumindest als Bewertungsmaß der linguistischen Features eines Textes dienen.

Mit *Listenability* soll im Folgenden das Hörverständnis beschrieben werden, also wie gut ein Text verständlich ist, wenn er angehört wird. Die Forschung zu Listenability jedoch ist sehr rar und das Messen und Bewerten von Listenability bleibt daher schwierig. Verschiedene Studien zur Listenability nutzen Readability Formeln [Chall and Dial, 1948, Denbow, 1975, Young, 1950], doch wurde deutlich, dass dies nur mit entsprechenden Anpassungen funktionieren kann [Chall and Dial, 1948, Dubay, 2007]. Da aussagekräftige Forschungsergebnisse zu den genauen Faktoren, die die Listenability beeinflussen, noch fehlen, blieb die genaue Art dieser »Anpassungen« jedoch lange unklar. Erst Kotani et al. [2014] stellten eine Methode vor, die Listenability unter Einbeziehung sowohl linguistischer Eigenschaften wie Sprechgeschwindigkeit und Wort- und Satzlänge⁶ als auch der Hörfertigkeiten der hörenden Person misst. Diese Methode erwies sich als effektiv, wurde jedoch bisher nur mit Transkripten und noch nicht mit automatischer Spracherkennung getestet. Des Weiteren ließen Kotani and Yoshimi [2017] die hörende Person ein Transkript des Audiomaterials erstellen und bewerteten anschließend die Listenability unter Einbeziehung der subjektiven Meinung der hörenden Person sowie der objektiven Auswertung des erstellten Transkriptes. Auch diese Methode erwies sich als effektiv und die Verlässlichkeit und Validität wurde von den Autoren als hoch bewertet. Mit einem reduzierten Einfluss des subjektiven Bewertungsteils wurde die Verlässlichkeit sogar noch erhöht.

Die Schwierigkeit bei der Bewertung von Listenability liegt vor allem darin festzustellen, welche Variablen eine Bewertungsformel enthalten muss. Messerklinger [2006] untersuchte auf der Basis von 41 Nachrichtentranskripten ver-

⁵<https://support.microsoft.com/en-us/office/get-your-document-s-readability-and-level-statistics-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2>, accessed 14.12.2020

⁶Alle verwendeten Eigenschaften waren Wort- und Satzlänge, mehrsilbige Worte, Sprechgeschwindigkeit, Schwierigkeit von Wörtern und phonologische Modifikationen

schiedene Faktoren wie Sprechgeschwindigkeit, Pausenlänge, Linguistische Komplexität, Akzent und Hintergrundwissen und konnte in seiner Studie nur für Hintergrundwissen, Vokabular und Störgeräusche einen deutlichen Einfluss auf die Listenability feststellen. Bereits 1943 fand Flesch heraus, dass »If material is put on the air rather than on the printed page, easy matter becomes easier but difficult matter more difficult. Radio will therefore magnify differentials in difficulty detected by the writer's formula.« Harwood [1955] unterstützt diese Erkenntnis: Auch er stellte fest, dass Material eher lesend verstanden wird als hörend, je schwieriger es wird. Ebenso konnte Dubay [2007] nach der Auswertung der Listenability von 132 in einfache Sprache umgeschriebenen automatischen Nachrichten eines Service-Telefons in Los Angeles County folgern, dass bessere Readability die Listenability erhöht. Chall and Dial [1948] legten 100 Erstsemestern 18 Radio-Skripte vor, die anhand der Flesch und der Dale-Chall Formeln in verschiedene Schwierigkeitsgrade eingeteilt worden waren. Auf diese Weise fanden sie Belege dafür, dass Hörverständnis schwieriger ist als Leseverständnis und empfahlen anschließend, Radionachrichten so einfach wie möglich zu formulieren. Auch wenn Young [1950] aus seiner Studie, bei der ein Text über ein komplexes Thema in vier verschiedenen Schwierigkeitsgraden⁷ formuliert und alle vier Versionen 128 Studenten vorgespielt wurden, folgerte, dass einfacher Stil nicht die alleinige Lösung für hohe Listenability sein könne, lässt sich doch ein deutlicher Trend der Ergebnisse hin zu einfachen Formulierungen erkennen. Dies wird auch durch die oben bereits genannten Erkenntnisse von Ortmann and Dipper [2019] unterstützt, dass oraler Diskurs einfachere Satzstrukturen enthält und weniger komplex ist.

2.3 Ratgeberliteratur zu Listenability

Die Ratgeberliteratur zum Thema »Schreiben fürs Hören« ergänzt die wissenschaftlichen Ergebnisse um Praxiserfahrungen. Journalist und ehemaliger Leiter der Henri-Nannen Journalistenschule Wolf Schneider und Chefredakteur und Autor Paul-Josef Raue empfehlen in ihrem Journalisten-Handbuch [Schneider and Raue, 2012] »einfache Sätze«, die häufigere Wiederholung des »handeln-

⁷anhand der Dale-Chall Formel

de[n] Subjekt[s]« und das Runden komplizierter Zahlen, um einen Text für Zuhörer verständlich zu formulieren. Sie fassen zusammen: »Für Texte zum Lesen und Texte zum Hören gelten dieselben Gesetze der Verständlichkeit [...]: die konkreten, die schlichtesten möglichen Wörter in schlanken, transparenten Sätzen.« Auch Jules [Horne, 2019b], Autorin für Radio und Theater, empfiehlt kurze Sätze, einfache Wörter, unkomplizierte Strukturen, bildhafte Beschreibungen sowie das Wiederholen der wichtigsten Punkte. Einen weiteren Schwerpunkt legt sie auf die Wichtigkeit des »Flows« und empfiehlt den gezielten Einsatz von Diskursmarkern, um den Fokus in eine bestimmte Richtung zu lenken und Strukturen zu erzeugen, die in schriftlichen Texten beispielsweise durch Überschriften, Absätze oder Listen zustanden kommen [Horne, 2019c]. McCormick [2013] weist Diskursmarkern sogar eine »key role« in Texten zu, »because [...] they help our creative and critical thoughts to flow more effectively.«

Zusammenfassend wird deutlich, wie wichtig es ist, die Fähigkeiten der wahrnehmenden Person in die Konzeption eines Textes mit einzubeziehen und den Text so einfach wie möglich zu formulieren. Dies gilt jedoch gleichermaßen für Listenability und Readability. Ein Punkt, der Listenability offenbar von Readability unterscheidet, ist der gezielt Einsatz von Diskursmarkern[Horne, 2019c, McCormick, 2013, Ortmann and Dipper, 2019], um bewusst Strukturen mit hoher Listenability zu kreieren. Kapitel 3 setzt sich daher mit dem Begriff *Diskursmarker* und der existierenden Forschung in diesem Feld auseinander.

3

Diskursmarker

Der Begriff *Diskursmarker* ist bis heute nicht klar definiert [Blühdorn et al., 2017]; »there is considerable variation in what might be labelled Discourse Markers.« [Fraser, 2009]. Auch der Begriff selbst variiert stark, u.a. wurden in der englischsprachigen Forschung *Cue Phrase*, *Discourse Connectives*, *Discourse Operators* und *Pragmatic Expressions* genutzt [Fraser, 2009]. Für einen detaillierten Überblick über »die Begriffsgeschichte [des Diskursmarkers] und die wichtigsten Inhalte der Diskursmarkerforschung« ist die Arbeit von Blühdorn et al. zu empfehlen. Abschnitt 3.1 dieses Kapitels beschäftigt sich mit verbreiteten Definitionen des Diskursmarkerbegriffs und definiert, was in dieser Arbeit darunter verstanden wird. In einem weiteren Schritt werden die unter diesem Begriff versammelten Worte in Abschnitt 3.2 in Bedeutungsgruppen eingeteilt, um eine tiefgehendere Analyse der Verwendung zu ermöglichen.

3.1 Begriff

Die Uneindeutigkeit des Diskursmarkerbegriffs begründet Fraser [2009] unter anderem mit den unterschiedlichen Zielen, die die Forschung im Bezug auf Diskursmarker in ihren Anfängen hatte. So beschäftigte sich Schiffarin [1987] vorrangig unter dem Gesichtspunkt der Gesprächsforschung mit ihnen. In diesen anfänglichen Forschungen wurden unter Diskursmarkern vor allem Worte verstanden, die ein Gespräch strukturieren, Verknüpfungen im Gespräch herstellen und für Kohärenz sorgen [Blühdorn et al., 2017]. Fraser [1988, 1990, 1999, 2009] gibt eine formalere, funktionale Definition des Diskursmarker-Begriffs,

und kommt damit Modellen aus der Computerlinguistik nahe [Blühdorn et al., 2017]. In dieser Richtung wird der Begriff heute vor allem verwendet, um Kohärenzrelationen anzuzeigen und ist nicht mehr nur auf Sprachhandlungen festgelegt. [Blühdorn et al., 2017].

Fraser geht davon aus, dass **DISCOURSE MARKERS** (DMs) eine Untergruppe der von ihm definierten *pragmatic markers* darstellen und daher nicht zur inhaltlichen Bedeutung des Satzes beitragen. Ihre Aufgabe ist es vielmehr, die Relation zwischen dem Diskurssegment, das den DM beinhaltet, und dem vorherigen Diskurssegment zu signalisieren. Um als DM gelten zu können, muss ein Ausdruck laut Fraser in der Sequenz <S1 DM S2> akzeptabel sein, wobei S1 und S2 Diskurssegmente darstellen. Es werden drei notwendige und hinreichende Bedingungen gegeben, die ein DM erfüllen muss:

1. Ein DM ist ein Inhaltswort
2. In einer Sequenz <S1 S2> muss ein DM, wenn er auftritt, ein Teil von S2 sein.
3. Ein DM trägt nicht zur semantischen Bedeutung der Sequenz bei, sondern signalisiert eine Relation zwischen S1 und S2

Nicht-definierend aber typisch für DMs ist, dass sie die Relation zwischen Diskurssegmenten nicht beeinflussen, sondern nur anzeigen. Würden sie fehlen, bliebe die Relation dieselbe. Des Weiteren sind DMs eine funktionale Klasse und setzen sich vor allem aus Wortgruppen wie Konjunktionen, Präpositionen und Adverbien, aber auch aus Verben und Nomen zusammen [Blühdorn et al., 2017, Fraser, 2009]. Eine großes Problem ist folglich die Mehrdeutigkeit. Ohne tiefgehende Analyse der betreffenden Diskurssequenz kann nicht eindeutig unterschieden werden, ob ein Wort an einer bestimmten Stelle ein DM ist oder sein Homograph. Auf den ersten Blick mögen hier Verfahren der *Word-Sense-Disambiguation* hilfreich erscheinen, doch ist dies ein separater Problembereich. Mit entsprechenden Algorithmen¹ kann zwar in den meisten Fällen bestimmt werden, ob das Wort *Bank* im Diskurszusammenhang eine Parkbank oder eine Bank zum Geldabheben meint. Die Phrase *as a result* hat jedoch in

¹<https://towardsdatascience.com/a-simple-word-sense-disambiguation-application-3ca645c56357>, accessed 15.12.2020

(1) She had watched the movie six times. **As a result**, she knew it by heart. (DM)

die gleiche Bedeutung wie in

(2) She knew the movie by heart **as a result** of watching it six times. (Adverb)

Da Fraser DM als eigene Wortgruppe definiert, kann ein Wort nicht gleichzeitig DM und beispielsweise Adverb sein. In welchem Fall ein potentieller DM nun aber die Funktion eines DM erfüllt und in welchem die seines Homographen aus einer anderen Wortgruppe, kann nicht mit für diese Arbeit vertretbarem Aufwand allgemeingültig bestimmt werden. In der empirischen Analyse werden folglich auch Homographen der DM mitgezählt.

Als eine weitere Untergruppe der *pragmatic markers* nennt Fraser die DISCOURSE STRUCTURE MARKERS (DSMs). Diese Marker tragen zur Organisation des laufenden Diskurses bei und sind laut Frasers Definition keine DMs. Die unklare Definition des Diskursmarker-Begriffs führt jedoch dazu, dass nicht mit Sicherheit gesagt werden kann, welche Wortgruppe die in Kapitel 2 diskutierte Literatur je genau meint. Ausgehend von den dort verwendeten Begriffen *Discourse Markers* [Horne, 2019c], *Connectives* [McCormick, 2013] und *Coordination* [Ortmann and Dipper, 2019] kann aber angenommen werden, dass ähnliche Strukturen gemeint sind: In allen Fällen werden mit den unterschiedlichen Begriffen Worte und Phrasen beschrieben, die Diskursrelationen anzeigen, Diskusssegmente miteinander verbinden und dem Hörer als Wegweiser im Text dienen sollen. Diese Eigenschaften würden auch die von Fraser definierten DSMs mit einschließen.

Eine frei verfügbare Quelle englischer *Discourse Connectives* ist die von Das et al. [2018] erstellte `en_dimlex.xml`² (EnDimLex). Diese Liste enthält 149 englische Connectives mit Informationen über syntaktische Kategorien, Diskurssemantik und die Verwendung als Nicht-Connective. Sie entstand durch das Zusammenfügen des annotierten Penn Discourse Treebank Corpus [Prasad et al., 2008], des RST Signalling Corpus [Das et al., 2015] und einer Liste mit Relations-Signalen [Biran and Rambow, 2011]. Connectives werden hier

²https://github.com/discourse-lab/en_dimlex/blob/master/en_dimlex.xml, accessed 15.12.2020

als Untergruppe der Diskursmarker betrachtet. Sie signalisieren eine Diskursbeziehung und setzen sich vorrangig aus Konjunktionen, Präpositionsphrasen, Adverbien, Nomen und Verben zusammen. Auch die Kriterien für die Aufnahme eines Wortes in die Sammlung ähneln den von Fraser definierten Bedingungen. Um als Discourse Connective im Sinne der EnDimLex-Liste zu gelten, muss ein Wort x ein Inhaltswort sein, für das gilt:

1. x kann nicht flektiert werden
2. x signalisiert eine zweiseitige Relation
3. Die Argumente dieser Relation sind abstrakte Objekte
4. Die Argumente können in einer Klausel-, Satz- oder Phrasenstruktur ausgedrückt werden

Des Weiteren muss x einen feststehenden, nicht modifizierbarer Ausdruck darstellen und darf nicht semantisch kombinierbar sein. Ersteres beschreibt Modifikationen von Phrasen wie *for this reason* zu *for this exact reason*, zweiteres Kombinationen von Connectives wie *particularly if*. Feststehende Phrasen wie *even if* jedoch wurden aufgenommen. EnDimLex enthält folglich Wörter, die von der Literatur unter verschiedenen Begriffen beschrieben und für die folgende Be trachtung aufgrund ihrer gemeinsamen funktionalen Eigenschaften unter dem Begriff *Diskursmarker* zusammengefasst werden. Diese Wörter sind feststehende Inhaltsworte, zeigen Relationen zwischen Diskurssegmenten an, sind nicht modifizierbar und tragen nicht zum Inhalt des Diskurses bei. Ihre Aufgabe in der Listenability ist, der hörenden Person Hinweise auf Diskursrelationen zu geben. Die in dieser Arbeit als Grundlage für die Analyse genutzten Diskursmarker sind daher die 149 Worte des EnDimLex.

3.2 Einteilung in Bedeutungsgruppen

Um nicht nur die gesamte Gruppe oder die einzelnen DM betrachten zu können, sondern auch Kategorien, deren DM je eine gleiche oder zumindest ähnliche Relation signalisieren, wurde beschlossen, die DM in Bedeutungsgruppen aufzuteilen. Wie bei der Unklarheit über den Begriff *Diskursmarker* an sich gibt

es auch bei ihrer Kategorisierung viele verschiedene Ansätze.

Fraser

Fraser teilt die von ihm definierten **DISCOURSE MARKERS** in drei Untergruppen auf und nennt die in den Beispielen jeweils zuerst aufgeführten DM *primary markers*, da sie bezeichnend für die Gruppe sind und je die umfassendste Bedeutung haben:

1. **CONTRASTIVE MARKERS** signalisieren einen Kontrast zwischen den Diskurssegmenten S1 und S2

but, alternatively, although, contrariwise, contrary to expectations, conversely, despite (this/that), even so, however, in spite of (this/that), in comparison, in contrast (to this/that), instead (of this/that), nevertheless, nonetheless, notwithstanding, on the other hand, on the contrary, rather (than this/that), regardless (of this/that), still, though, whereas, yet, ...

2. **ELABORATIVE MARKERS** signalisieren eine genauere Ausführung im folgenden Segment

and, above all, after all, also, alternatively, analogously, besides, by the same token, correspondingly, equally, for example, for instance, further (more), in addition, in other words, in particular, likewise, more accurately, more importantly, more precisely, more to the point, moreover, on that basis, on top of it all, or, otherwise, rather, similarly, ...

3. **INFERENTIAL MARKERS** signalisieren, dass S2 aus S1 gefolgert werden kann

so, all things considered, as a conclusion, as a consequence (of this/that), as a result (of this/that), because (of this/that), consequently, for this/that reason, hence, it follows that, accordingly, in this/that/any case, on this/that condition, on these/those grounds, then, therefore, thus, ...

Und auch wenn Fraser selbst die von ihm definierten DISCOURSE STRUCTURE MARKERS klar von den DMs abgrenzt, sind sie doch für unsere Analyse bedeutsam. DSMs teilt Fraser in drei weitere Untergruppen auf:

1. DISCOURSE MANAGEMENT MARKERS beziehen sich auf den gesamten Diskurs

in summary, I add, ...

2. TOPIC ORIENTATION MARKERS beziehen sich auf Themen des Diskurses

Returning to, continuing, ...

3. ATTENTION MARKERS lenken die Aufmerksamkeit in eine bestimmte Richtung

look, now, ...

Fraser kann nach dieser Aufteilung keinen DM finden, der in zwei Klassen fiele.

McCormick

McCormick [2013] nennt neun Hauptfunktionen für die von ihm beschriebenen CONNECTIVES:

1. Wo befindet sich etwas in Bezug auf etwas anderes

before, near, next, there, where, ...

2. Wann findet etwas statt

after, again, at last, earlier, next, once, whenever, ...

3. Zwei Ideen vergleichen und Ähnlichkeiten ausdrücken

additionally, again, equally, similarly, not to mention, ...

4. Zwei Ideen in Kontrast setzen und Unterschiede herausstellen

and yet, although, besides, but, despite, even though, rather, while, yet, ...

5. Zusätzliche oder unterstützende Ideen ausdrücken

additionally, again, also, and, as well, besides, equally, too, ...

6. Andeuten, dass etwas bereits betrachtet wurde

clearly, more particularly, generally, to be more precise, ...

7. Logische (Reihen-)Folge ausdrücken

accordingly, because, due to, hence, if since, therefore, thus, unless, while, ...

8. Ein Beispiel geben oder illustrieren

especially, for example, for instance, in fact, with regard to, ...

9. Zusammenfassung

after all, finally, hence, in fact, overall, that is, to sum up, usually, ...

Bei dieser Art der Aufteilung können Funktionen sich überlappen. Beispielsweise kann der DM *again* sowohl eine zeitliche Relation anzeigen (Funktion 2), Ähnlichkeiten zwischen zwei Ideen signalisieren (Funktion 3) als auch auf zusätzliche Ideen hinweisen (Funktion 5).

Horne

Horne [2019c] teilt DISCOURSE MARKERS in folgende drei Kategorien auf:

1. SEQUENCES

as an aside, others think, ...

2. ELABORATION

in other words, to paraphrase, ...

3. SUMMING UP

basically, in a nutshell, in short, ...

EnDimLex

Das et al. [2018] teilen die EnDimLex-Liste in folgende vier Bedeutungskategorien ein:

1. COMPARISON

although, but, in contrast, still, while, yet, ...

2. CONTINGENCY

accordingly, because, for, given, in case, whatever, ...

3. EXPANSION

also, and, as well, besides, finally, instead, in short, particularly, rather, ...

4. TEMPORAL

afterward, as, before, meanwhile, next, thereafter, simultaneously, ...

Diese vier Hauptbedeutungen werden in insgesamt 45 weitere Untergruppen aufgesplittet, die jedoch für unsere Zwecke zu detailliert sind. Auch bei dieser Aufteilung gibt es einige Überschneidungen, die für die Analyse beachtet werden müssten.

Vergleich

| EnDimLex | Fraser | McCormick | Horne |
|-------------|--|--|----------------------------|
| Comparison | Contrast | Differences, Similarities Logical Sequences | Sequence |
| Contingency | Inferential | Logical Sequence, Summarize | Sequence Summing up |
| Expansion | Elaborative | Additional Ideas, Examples, Summarize | Elaboration, Summing up |
| Temporal | Discourse Management, Topic Orientation | Logical Sequence, When | Sequence |

Tabelle 3.1: Darstellung der sich je ähnlichen DM-Klassen der verschiedenen Ansätze

All diese Ansätze lassen sich, wie in Tabelle 3.1 gezeigt, auf einen gemeinsamen Nenner bringen. Einzig die von McCormick vorgestellte Kategorie »Wo befindet sich etwas in Bezug auf etwas anderes« und die von Fraser aufgestellte Klasse der ATTENTION DSMs fallen aus dem Rahmen (wobei die ATTENTION DSMs ohnehin von Fraser nicht als DMs definiert werden) und können nicht eingeordnet werden.

Da sich alle anderen Ansätze gut in die vier Kategorien des EnDimLex einordnen lassen, wird im Folgenden die Einteilung des EnDimLex verwendet. Ein weiterer großer Vorteil dieser Einteilung ist, dass EnDimLex bereits Häufigkeitswerte für jedes Wort in jeder Kategorie enthält. Diese Häufigkeitswerte beziehen sich auf die annotierten Corpora, aus denen EnDimLex erstellt wurde [Das et al., 2018]. Mit Hilfe des `MarkerTypes.py`³ Scripts wurden Wörter, die in mehre-

³<https://git.webis.de/code-teaching/theses/thesis-sacher/-/blob/master/scripts/listenability-tools/MarkerTypes.py>

ren Kategorien erscheinen, für die Analyse in der Kategorie eingeordnet, in der sie am häufigsten vorkamen. Anschließend wurde mit `Ambiguity.py`⁴ für jedes Wort eine Error Rate berechnet, die angibt, wie groß die Fehlerwahrscheinlichkeit ist, wenn dieses Wort unter der ausgewählten Kategorie betrachtet wird. Es ist anzumerken, dass diese Error Rate nicht den Fehler mit einbezieht, der durch die Mitzählung der Homographen eines jeden Wortes entsteht. Tabelle A.1 im Anhang zeigt eine vollständige Liste aller analysierten Worte, eingeteilt in die vier Kategorien **COMPARISON**, **CONTINGENCY**, **EXPANSION** und **TEMPORAL**, inklusive der jeweiligen Error Rate.

⁴<https://git.webis.de/code-teaching/theses/thesis-sacher/-/blob/master/scripts/listenability-tools/Ambiguity.py>

4

Einteilung der Textdaten in Textsorten

In diesem Kapitel wird auf die Einteilung der zur Analyse verwendeten Textdaten in Kategorien bzw. *Textsorten* eingegangen. Es gibt dabei drei Arten der Unterscheidung: Abschnitt 4.1 beschreibt die Einteilung der Texte nach Diskursart, Abschnitt 4.2 die Einteilung nach Genre und Abschnitt 4.3 die Einteilung nach Konversationsart. Jede dieser Arten der Einteilung enthält verschiedene Kategorien, in welche die verwendeten Texte eingeteilt wurden, und jede dieser Kategorien stellt eine eigene Textsorte dar.

4.1 Diskursarten

Mit dem Begriff *Diskursarten* werden ORAL-AKUSTISCHER und LITERAT-SCHRIFTLICHER Diskurs beschrieben. Dabei sind *oral* und *literat* Konzepte, *akustisch* und *schriftlich* hingegen Medien [Ortmann and Dipper, 2019]. Konzepte und Medien können, wie in Tabelle 2.1 zu sehen, beliebig kombiniert werden, doch werden in der folgenden Untersuchung nur die oben genannten Kombinationen ORAL-AKUSTISCH und LITERAT-SCHRIFTLICH unterschieden. Sie sind von besonderem Interesse, da betrachtet werden soll, wie Texte, die für ihr jeweiliges Medium konzipiert wurden, Diskursmarker nutzen. Ein Vergleich der Diskursmarkernutzung beider Textsorten soll zeigen, ob Diskursmarker für das akustische Medium anders genutzt werden als für das schriftliche. Hierfür werden die vier in Kapitel 5 beschriebenen Corpora als Datengrundlage genutzt. Zwei davon fallen in die ORAL-AKUSTISCHE Kategorie, zwei in die LITERAT-SCHRIFTLICHE.

4.2 Genres

Der Begriff *Genre* beschreibt die Einteilung der Texte anhand ihrer Gattung¹. In dieser Arbeit werden nur die Genres der ORAL-AKUSTISCHEN Texte betrachtet. Folgende Genres werden in dieser Arbeit unterschieden und in der jeweiligen Corpusbeschreibung näher ausgeführt: NEWS, DISCUSSION, SCIENCE/EDUCATION, DOCUMENTARY und PRESENTATION. Als Datengrundlage werden die beiden in den Unterabschnitten 5.1.3.2 und 5.1.3.3 beschriebenen ORAL-AKUSTISCHEN Corpora genutzt, von denen einer in die Kategorie PRESENTATION fällt und der andere in die restlichen Genres aufgeteilt wird.

4.3 Konversationsarten

Der Begriff *Konversationsarten* beschreibt im Folgenden Variationen in der Form des akustischen Diskurses. Es wird hier unterschieden zwischen DIALOG, MONOLOG, KOOPERATIVEM MONOLOG und REDEN. Als DIALOG werden Texte eingeordnet, in denen vorrangig zwei oder mehr Personen miteinander sprechen, beispielsweise in einer Diskussion oder einem Interview. Spricht vorrangig eine einzelne Person mit sich selbst oder an die Zuhörer gewandt und geschieht dies nicht nach einem Skript, werden diese Texte als MONOLOG eingeordnet. Eine Sonderform des MONOLOGS stellt die in dieser Arbeit eingeführte Kategorie des KOOPERATIVEN MONOLOGS dar. In diese Kategorie fallen Texte, wenn zwar zwei oder mehr Personen sprechen, sie jedoch nicht direkt miteinander reden. Dies geschieht beispielsweise in Dokumentationen, bei denen Reporter von anderen übernehmen, um über einen separaten Teil des Themas zu berichten. Alle Beteiligten sprechen zwar koordiniert zum gleichen Thema, um einen vollständigen Bericht zu erschaffen, aber es ist keine Interaktion miteinander erkennbar. Als eine weitere Sonderform des MONOLOGS wird die REDE betrachtet. Hier spricht eine Person zu einem Publikum nach einem vorab konzipierten Skript. Was die REDE abhebt vom MONOLOG, ist die Interaktion mit dem Publikum – »eine Wechselwirkung, wie der Schreiber sie sonst nie erlebt.« [Schneider, 2011]. Als Datengrundlage werden wieder die beiden in den Unterabschnitten 5.1.3.2 und 5.1.3.3 beschrie-

¹<https://www.duden.de/rechtschreibung/Genre>, accessed 28.12.2020

benen ORAL-AKUSTISCHEN Corpora genutzt, von denen einer in die Kategorie REDE fällt und der andere aufgeteilt wird in DIALOG, MONOLOG und KOOPERATIVER MONOLOG.

5

Corpora

In diesem Kapitel wird ein Überblick über die zur Analyse verwendeten Corpora gegeben und Anforderungen an die Daten sowie Probleme bei der Datenfindung werden erläutert. Da für einen Vergleich sowohl akustische als auch schriftliche Corpora benötigt wurden, ist dieses Kapitel aufgeteilt in die entsprechenden Unterkapitel 5.1 und 5.2.

Die Pfade zu den verwendeten Python-Scripten sind relative Dateipfade im beinhaltenden Repository¹.

5.1 Akustische Corpora

5.1.1 Anforderungen

Um akustischen Diskurs zu untersuchen und mit schriftlichem Diskurs zu vergleichen, wird eine schriftliche Form des akustischen Diskurses benötigt. Diese steht mit Transkripten zur Verfügung, die entweder manuell oder automatisch erstellt werden. Die Art der Erstellung hat großen Einfluss auf die Qualität des Transkriptes, aber auch auf die mögliche Größe des Datensatzes. Manuell erstellte Transkripte sind oft genauer, das bedeutet, sie enthalten weniger Wortfehler und akkuratere Interpunktions, ihre Erstellung benötigt jedoch wesentlich mehr Zeit als die automatische Erstellung durch Algorithmen und beschränkt somit die Größe des Datensatzes. Eine große Menge an Daten ist allerdings für einen aussagekräftigen Vergleich unerlässlich. Der Themenbereich der Daten

¹<https://git.webis.de/code-teaching/theses/thesis-sacher/-/tree/master/scripts>

wurde zudem auf Nachrichten eingegrenzt. Um den besten Kompromiss zu finden, wurden vorab folgende Kriterien festgelegt:

- Kostenlose
- Groß
- Qualitativ hochwertige, von Experten erstellte Daten
- Mit dem entsprechenden schriftlichen Nachrichtenartikel verlinkt
- Sowohl Nachrichten als auch Editorials
- Daten aus vielen verschiedenen Quellen
- Inklusive Keywords, die das Thema angeben

Einen kostenlosen Corpus mit Nachrichten-Transkripten zu finden, erwies sich jedoch als nahezu unmöglich; geschweige denn einen Corpus, der all die oben genannten Kriterien erfüllt. Diese wurden also folgendermaßen angepasst: An erster Stelle steht der kostenlose Zugang zu den Daten, gefolgt von einer guten Abwägung zwischen Größe und Qualität. Da kostenlose Nachrichten-Corpora schwer zu finden waren, sollte das Thema zumindest überwiegend sachlich sein. Verlinkungen zu entsprechenden schriftlichen Nachrichten wurden gestrichen. Eine vollständige Liste in Frage kommender Datensätze findet sich in Tabelle A.2, auch wenn letztendlich nur drei dieser acht Corpora tatsächlich eine ausreichende Menge der Kriterien aufwiesen und näher betrachtet wurden. Außerdem wurde deutlich, dass Interpunktions für den gewünschten Anwendungsfall von Bedeutung ist, da Eigenschaften wie die Positionierung von Diskursmarkern im Satz sich nur prüfen lassen, wenn die Referenz-Daten eine erkennbare Satzstruktur aufweisen. Auf dieses besondere Problem wird in Unterabschnitt 5.1.2.1 eingegangen.

5.1.2 Probleme automatisch erstellter Transkripte

5.1.2.1 Problem der fehlenden Interpunktions

Gesprochene, spontane Sprache enthält keine Interpunktions, der Gedankenfluss wird hier durch Stimmveränderungen und Pausen strukturiert. Bei vorgelesenen Texten liegt es am Sprecher, die Interpunktions entsprechend zu interpretieren

und darzustellen. Beides führt zu fehlender oder fehlerhafter Interpunktions in automatisch erstellten Audio-Transkripten, da eine zweisekündige Pause sowohl einen Punkt als auch ein Komma oder nur eine Pause signalisieren kann. Und obwohl eine Vielzahl von Bibliotheken existiert, die *Sentence Segmentation*, also das korrekte Aufteilen des Textes in Satzfragmente, beherrschen (z.B. NLTK² und Stanford CoreNLP³), funktionieren diese am besten auf Texten, die korrekte Groß- und Kleinschreibung sowie Interpunktions enthalten [Bedapudi, 2018]. Das Einfügen von Interpunktions basierend auf dem Diskurskontext ist nach wie vor eine Herausforderung.

Im Laufe dieser Arbeit wurden drei Verfahren getestet, Interpunktions zu automatisch erstellten Transkripten hinzuzufügen. Um die Ergebnisse bewerten zu können, wurde zunächst in einige ausgewählte Transkripte die Interpunktions manuell eingefügt. Anschließend wurde jedes der unten beschriebenen Verfahren auf die ursprünglichen Transkripte angewendet und die Ergebnisse mit der zuvor von uns manuell bearbeiteten Version verglichen. Diese Vorgehensweise wurde für den in Unterunterabschnitt 5.1.3.1 beschriebenen RadioTalk Corpus sowie den in Unterunterabschnitt 5.1.3.3 beschriebenen TED-LIUM 3 Corpus angewendet, da beide keine Interpunktions aufweisen. In keinem Fall konnte jedoch ein ausreichend gutes Ergebnis erzielt werden.

Pausenbasierte Interpunktions

In einem ersten Versuch wurde ein Punkt an jeder im Transkript markierten Pause eingefügt. Dies führte nicht zu ausreichend guten Ergebnissen, da Pausen, wie oben beschrieben, kein sicherer Indikator für den Abschluss eines Gedankengangs sind. Es ist ebenso möglich, dass die sprechende Person an dieser Stelle unterbrochen oder abgelenkt wurde, kurz überlegen musste oder einfach einen Schluck Tee getrunken hat.

²<https://www.nltk.org/api/nltk.tokenize.html>, accessed 16.12.2020

³<https://stanfordnlp.github.io/CoreNLP/ssplit.html>, accessed 16.12.2020

Interpunktions mit Deepsegment

Im nächsten Schritt wurde das Python Modul Deepsegment⁴ angewendet, das mit einer Million Beispielen trainiert wurde. Bei vorhandener Interpunktions kann dieses Modul laut Angaben des Autors [Bedapudi, 2018] mit einer Wahrscheinlichkeit von 97% den Text in korrekte Satzfragmente aufteilen. Bei fehlender Interpunktions kommt es auf eine Genauigkeit von 53%. In keinem der getesteten Fällen wurden mit diesem Modul ausreichend gute Ergebnisse erzielt, wie in Tabelle 5.1 und Tabelle 5.2 anhand von Stichproben deutlich zu sehen ist.

BERT Masked Language Model

Ein letzter Versuch wurde mit dem *pretrained BERT masked language model*⁵ für Python durchgeführt. BERT steht für Bidirectional Encoder Representations from Transformers [Devlin et al., 2018] und ist ein bidirektionales Sprachmodell, das u.a. mit einem *Masked Language Model* (MLM) trainiert wird. Hierbei werden zufällig einige der Tokens maskiert. Das Ziel für BERT ist dann, das ursprüngliche Wort nur anhand des Kontexts vorauszusagen. Für dieses Verfahren wurden alle im Transkript markierten Pausen maskiert und anschließend mit BERT vorausgesagt. Jedoch brachte auch dieses Verfahren für keinen der Corpora ein ausreichendes Ergebnis, wie wieder in Tabelle 5.1 und Tabelle 5.2 gut zu sehen ist. Dies liegt zum Teil daran, dass, wie bereits oben beschrieben, nicht jede Pause ein Satzzeichen garantiert und BERT außerdem auch ganze Wörter voraussagt. Mögliche Satz trennungen an nicht maskierten Stellen im Text konnten auf diese Weise ebenfalls nicht vorausgesagt werden.

5.1.2.2 Probleme der Bewertung mit der Word Error Rate

Die *Word Error Rate* (WER) ist ein Maß zur Bewertung von automatischer Spracherkennung und wird in dieser Arbeit bei der Bewertung der Transkript-Qualität als unterstützender Faktor hinzugezogen. Die WER wird wie folgt bestimmt⁶:

⁴<https://github.com/notAI-tech/deepsegment>, accessed 16.12.2020

⁵<https://github.com/huggingface/transformers>, accessed 16.12.2020

⁶<https://www.rev.ai/blog/how-to-calculate-word-error-rate/>, accessed: 16.12.2020

$$WER = \frac{S + I + D}{N}$$

N = Anzahl der Wörter im Original

(1) What have the Romans ever done to us?

S = Anzahl der Substitutions

(2) What have the Omens ever done to us?

I = Anzahl der Insertions:

(3) What have the Romans Omens ever done to us?

D = Anzahl der Deletions:

(4) What have the _____ ever done to us?

Je höher die WER, desto ungenauer demnach das Transkript. Es steht jedoch zur Diskussion, ob die WER die Qualität der Spracherkennung tatsächlich angemessen bewertet. Unter anderem wird argumentiert, dass auf diese Weise alle Worte den gleichen Wert haben, wichtige Wörter werden genauso bewertet wie beispielsweise Diskursmarker, die nichts zum Inhalt beitragen. Zudem würde beispielsweise

(1) »What have the Omens ever done to us?«

dieselbe WER erzielen wie

(2) »What have the ever done to us?«

wobei Beispiel (1) für menschliche Leser wesentlich nützlicher wäre als Beispiel (2)⁷. Ein weiteres Problem sind Unterschiede in der Normalisierung. Es kann beispielsweise sein, dass Abkürzungen und Interpunktionszeichen nicht mit in die angegebene WER eingerechnet wurden⁸. Zwar existiert laut Favre et al. [2013] sehr

⁷<https://medium.com/descript/challenges-in-measuring-automatic-transcription-accuracy-f322bf5994f>, accessed 16.12.2020

⁸<https://deepgram.com/blog/the-trouble-with-wer/>, accessed: 17.12.2020

wahrscheinlich keine bessere Methode, doch wird die WER in dieser Arbeit aufgrund der genannten Faktoren mit einer gewissen Vorsicht zur Bewertung der Transkript-Qualität verwendet.

5.1.3 Corpora

5.1.3.1 RadioTalk Corpus

Überblick

Der RadioTalk Corpus [Beeferman et al., 2019] schien anfangs den gewünschten Kriterien am nächsten zu kommen. Er enthält Transkripte von Talk-Radio Sendungen aus den USA zwischen Oktober 2018 und März 2019, die mit Hilfe des Kaldi Speech Recognition Toolkits [Povey et al., 2011] erstellt wurden. Die Daten bestehen aus ca. 2.8 Milliarden Wörtern von 284 000 Stunden Audioaufnahmen einer großen Bandbreite von Sprechern in sowohl freier als auch geskripteter Rede. In den Daten enthalten sind sowohl Metadaten über Sprecher, Programm und Sender als auch die Transkripte der Audio-Snippets. Diese wurden basierend auf Sprecherwechseln und Pausen in »Snippets« geteilt (s. Abbildung A.1). Sie weisen eine Word Error Rate von 13.1% auf und beinhalten keine Interpunktionszeichen, was letztendlich dazu führte, dass der Corpus nicht verwendet werden konnte.

Corpusanalyse

Der Corpus besteht aus einer 50 GB großen JSON Datei im JSON Lines Format, beispielhaft zu sehen in Abbildung 5.1. Mit dem `SimpleCorpusParser.py`⁹ wurde die Datei geparsed und alle gefundenen Shownamen wurden gespeichert, um zunächst einen Überblick über die Datenmenge und die vertretenen Genres zu erhalten. Da von einigen Shows mehrere Episoden enthalten sind, bestand die resultierende Liste noch aus über 200 000 Einträgen. Mit dem Script `ShowMerger.py`¹⁰ wurden mehrfach vorkommende Shows gezählt und eine neue Liste mit knapp 1000 Shows erstellt, die jede Show und deren jeweiligen Episo-

⁹`RadioTalk/SimpleCorpusParser.py`

¹⁰`RadioTalk>ShowMerger.py`

```
{"content": "Yeah they look away they they literally he ... on the system they\nuh there's two gangs of southerners and the there's the Mexican", "callsign":\n"KABC", "city": "Los Angeles", "state": "CA", "show_name": "The Drive Home With\nJillian Barberie & John Phillips", "signature": "dcbd6cd5",\n"studio_or_telephone": "T", "guessed_gender": "M", "segment_start_time":\n1540944205.0, "segment_end_time": 1540944215.7, "speaker_id": "S0",\n"audio_chunk_id": "2018-10-31/KABC/00_03_25/0"}\n{"content": "And by the bulldogs they won the entire prison system then the\nwhite barely fit into blocks on the road show it's tribalism at the at a\nlow", "callsign": "KABC", "city": "Los Angeles", "state": "CA", "show_name":\n"The Drive Home With Jillian Barberie & John Phillips", "signature":\n"6b4c53ee", "studio_or_telephone": "T", "guessed_gender": "M",\n"segment_start_time": 1540944215.7, "segment_end_time": 1540944225.05,\n"speaker_id": "S0", "audio_chunk_id": "2018-10-31/KABC/00_03_25/1"}
```

Abbildung 5.1: Beginn eines RadioTalk Transkripts

denzähler nur noch einmal enthält. Da die Daten keinerlei Genre- oder Inhalts-Informationen enthalten und der Begriff *Talk-Radio* nur besagt, dass die Show einen hohen Wortanteil hat¹¹, war nach wie vor fraglich, ob und wenn ja wie viele geeignete Shows in den Daten zu finden sein würden. 1000 Shows sind bei einer Recherchezeit von ca. 1-8 Minuten pro Show, also ca. 30 bis 35 Stunden Arbeitszeit, eine zu große Menge, um in angemessener Zeit manuell recherchierbar zu sein. Daher wurde mit einem einfachen Suchfilter geprüft, ob geeignete nachrichtenähnliche Shows im Datensatz enthalten waren: Die Liste der Shownamen wurde nach den Stichworten *News*, *Daily*, *Morning* und *Report* durchsucht. Zusätzlich wurden einige zufällig ausgewählte Shows recherchiert. Die resultierende Liste mit Genre-Einordnungen und Einschätzungen zur Verwendbarkeit findet sich in Tabelle A.3. Letztendlich wurden 49 Shows als verwendbar eingestuft. Der Inhalt dieser Programme sollte nun zu kohärenten Texten zusammengefügt werden. Anhand der enthaltenen `audio_chunk_id` wurden die Snippets mit dem `SnippetMerger.py`¹² erfolgreich zu zusammenhängenden Texten zusammengesetzt, jedoch fiel nun auf, dass sie keine Interpunktionszeichen enthielten. Dieser Post-Processing Schritt war für die Autoren des Corpus nicht notwendig gewesen: »For our own linguistic analyses of the corpus, we have usually treated snippets as “noisy” sentences, good enough for topic analysis but admittedly error-prone if you rely on the parse of the sentence in some

¹¹<https://de.wikipedia.org/wiki/Talkradio>, accessed 17.12.2020

¹²[RadioTalk/SnippetMerger.py](#)

way in your downstream application.«¹⁸. Mit den in Unterunterabschnitt 5.1.2.1 beschriebenen Verfahren wurde versucht, Interpunktionszeichen einzufügen, doch war hierfür die Qualität der Transkripte zu schlecht. Ein Beispielhaftes Ergebnis ist in Tabelle 5.1 aufgeführt. Letztendlich wurde der Corpus für diese Arbeit nicht weiter verwendet.

| Manuelle Interpunktionszeichen | | |
|--|---|--|
| Pausenbasiert | DeepSegment | BERT |
| It's broke preserving manage their wealth learn more for that whether you dot com slash wealth fidelity brokerage services Pelosi And garbled a meal for a service delivering, fresh gourmet dinner kitchen table ready in Fifteen minutes with just one pan, gobbled does the prep work, you spend less time shopping and more time with family, more and more global dot com slash BBC. And we would like to thank the local businesses which provided commentary trees for volunteers that during a recent membership Dr including airs Mandy bakery for ... and pastries things cheesecake for their fresh cheese cakes ... next visit please thank them for supporting your local public radio station. You can find links to those and other food donors At our website KLW that oh gee [noise]. | It's broke preserving manage their wealth learn more for that whether you dot com slash wealth fidelity brokerage services Pelosi And garbled a meal for a service delivering fresh gourmet dinner kitchen table ready in Fifteen minutes with just one pan gobbled does the prep work you spend less time shopping and more time with family more and more global dot com slash BBC. And we would like to thank the local businesses which provided commentary trees for volunteers that during a recent membership Dr including airs Mandy bakery for ... and pastries things cheesecake for their fresh cheese cakes ... next visit please thank them for supporting your local public radio station. You can find links to those and other food donors At our website KLW that oh gee [noise] | It's broke preserving manage their wealth learn more for that whether you dot com slash wealth fidelity brokerage services Pelosi bread And garbled a meal for a service delivering fresh gourmet dinner kitchen table ready in Fifteen minutes with just one pan gobbled does the prep work you spend less time shopping and more time with family more and more global dot com slash BBC ##s And we would like to thank the local businesses which provided commentary trees for volunteers that during a recent membership Dr including airs Mandy bakery for ... and pastries things cheesecake for their fresh cheese cakes ... next visit please thank them for supporting your local public radio station if You can find links to those and other food donors like At our website KLW that oh gee [noise]. |

Tabelle 5.1: Veranschaulichung der hohen WER des RadioTalk Corpus und der unzureichenden Qualität der eingefügten Interpunktionszeichen. Rot markiert sind die vom jeweiligen Verfahren eingefügten Symbole.

¹⁸private E-Mail Konversation mit Doug Beeferman vom 15.6.2020, s. Abbildung A.1

5.1.3.2 Spotify Podcast Corpus

Überblick

Der Spotify Podcast Corpus [Clifton et al., 2020] enthält mehr als 100 000 Spotify Podcast Episoden mit fast 60 000 Stunden transkribiertem Audiomaterial. Abgedeckt werden eine große Bandbreite an Produzenten, Genres und Konversationsarten, sowie geskripteter und improvisierter Diskurs. Die Episoden wurden nach Sprache, Länge und Sprachanteil gefiltert. Es sind nur englischsprachige Podcasts enthalten und nur professionell produzierte Shows sind länger als 90 Minuten. Außerdem wurden Episoden herausgefiltert, die weniger als 50% Sprachanteil haben [Clifton et al., 2020]. Die Transkripte wurden automatisch mit Hilfe von Googles Cloud Speech-to-Text API¹⁴ erstellt und enthalten Interpunktions sowie Groß- und Kleinschreibung. Trotz der WER von 18.1% war die Qualität der Transkripte ausreichend für unsere Analysen.

Corpusanalyse

Die Transkripte des Spotify Podcast Corpus wurden als JSON Dateien zur Verfügung gestellt. Für jede Episode existiert eine eigene Datei, die das vollständige Transkript enthält, beispielhaft in Abbildung 5.2 zu sehen.

```
{"results": [{"alternatives": [{"transcript": "Hello and welcome to the\nvolunteer firefighter podcast where we listen in to a group of rural\nfirefighters as they give their opinions on the challenges. They face both on\nand off the fire ground. We release a new episode every week. So please hit\nthat subscribe button leave us a rating and share this with your fire family\nand friends now onto this week's episode. Where as always we ask the question.\nAre you DTF?", "confidence": 0.9094750285148621, "words": [{"startTime": "1.\n400s", "endTime": "2s", "word": "Hello"}, {"startTime": "2s", "endTime": "2.\n300s", "word": "and"}, {"startTime": "2.300s", "endTime": "2.900s", "word": "welcome"}, {"startTime": "2.900s", "endTime": "3s", "word": "to"}, {"startTime": "3s", "endTime": "3.200s", "word": "the"}, {"startTime": "3.\n200s", "endTime": "3.900s", "word": "volunteer"}, {"startTime": "3.900s", "endTime": "4.600s", "word": "firefighter"}, {"startTime": "4.600s", "endTime": "5.300s", "word": "podcast"}, {"startTime": "5.700s", "endTime": "5.900s", "word": "where"}, {"startTime": "5.900s", "endTime": "6s", "word": "we"}]}]}
```

Abbildung 5.2: Beginn eines Spotify Podcast-Transkripts

¹⁴<https://cloud.google.com/speech-to-text/docs/video-model>, accessed 18.12.2020

Laut der anfangs aufgestellten Kriterien mussten für die Analyse Podcasts herausgesucht werden, die bestenfalls Nachrichtensendungen waren. Zwar enthalten die Metadaten der Podcasts unter anderem eine von den Produzenten angegebene Beschreibung, doch ist diese oft sehr ungenau [Clifton et al., 2020]. Die Metadaten stellen aber zusätzlich für jede Show einen RSS Link zur Verfügung, welcher unter dem Tag <itunes:category> ein Genre enthält. Mit den Scripten `AnalyzePodcastData.py`¹⁵ und `GenreList.py`¹⁶ wurde daher zunächst eine Liste der vorhandenen Genres erstellt, die anschließend manuell nach relevanten Genres durchsucht wurde. Die Genres *Business News*, *Daily News*, *News*, *Sports News* und *Tech News* wurden dabei als relevant markiert. Da für diese Genres insgesamt nur 100 Shows gefunden wurden, von denen sich nach weiterer Recherche letztendlich nur 52 als relevant erwiesen, wurde nach weiteren Genres mit »sachlichem« Stil gesucht. Als größtenteils sachlich, gut recherchiert und geskriptet wurden die Genres *Documentary*, *History*, *Science* und *True Crime* eingeschätzt. Zusammen mit den bereits gefundenen News-Shows ergab sich die in Tabelle A.4 zu sehende Auflistung. 140 relevante Podcast Shows mit insgesamt 2 782 Episoden und rund 17 Millionen Wörtern konnten zusammenge stellt werden. Mit dem `ShowContent.py`¹⁷ Script wurden nun die Inhalte der entsprechenden Episoden in je eine TXT-Datei geschrieben und zur weiteren Verwendung abgespeichert.

Um im Verlauf der Arbeit einen Vergleich zwischen verschiedenen Genres aufzustellen zu können, wurden die relevanten Podcasts in vier Genres eingeteilt:

1. NEWS

Der Fokus liegt vor allem auf der Übermittlung von Nachrichten.

2. DISCUSSION

Shows, die aus Diskussionen und Meinungsaustausch bestehen.

¹⁵Spotify/AnalyzePodcastData.py

¹⁶Spotify/GenreList.py

¹⁷Spotify>ShowContent.py

3. SCIENCE/EDUCATION

Shows, die allgemeines oder spezielles Wissen vermitteln.

4. DOCUMENTARY

Geskriptete, gut recherchierte Dokumentationen zu einem bestimmten Thema, meist aufwendig produziert.

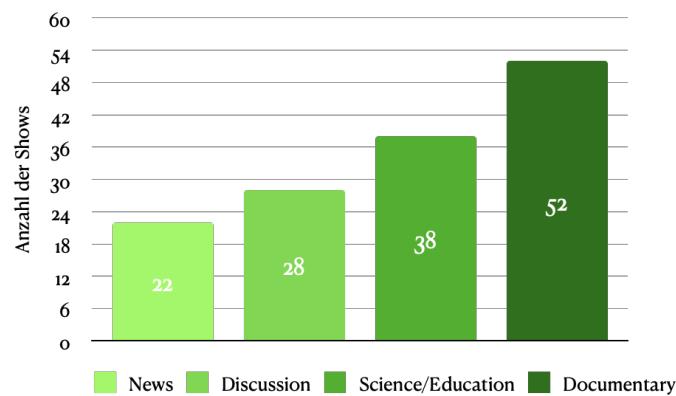


Abbildung 5.3: Verteilung der Shows nach Genre

Basierend auf Genre und Stich-Hörproben in jeden Podcast wurden die Shows außerdem in die drei Konversationsarten DIALOG, MONOLOG und KOOPERATIVER MONOLOG eingeteilt, wie unter Abschnitt 4.3 beschriebenen.

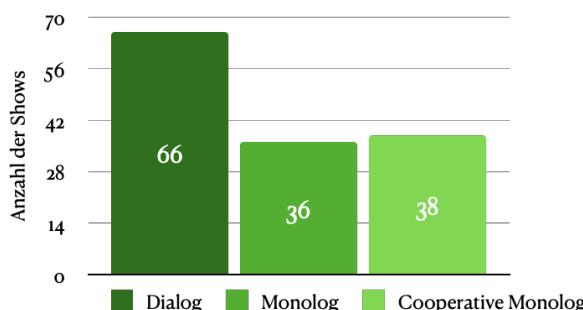


Abbildung 5.4: Verteilung der Shows nach Konversationsart

5.1.3.3 TED-LIUM 3 Corpus

Überblick

Der TED-LIUM 3 Corpus [Hernandez et al., 2018] besteht aus 1983 Reden mit insgesamt rund vier Millionen Wörtern. Die Transkripte wurden automatisch mit Hilfe des Kaldi-Toolkits [Povey et al., 2011] erstellt und weisen eine WER von 6.7% auf. Sie enthalten keine Interpunktions- und keine Genre-Informationen.

Corpusanalyse

Die Transkripte wurden im STM Format zur Verfügung gestellt, ein Beispiel ist in Abbildung 5.5 zu sehen.

```
MagnusLarsson_2009G 1 MagnusLarsson_2009G 12.7 19.19 <NA> <unk> funny to be at a  
conference dedicated to things not seen <unk> and present my proposal to build  
a <unk>  
MagnusLarsson_2009G 1 MagnusLarsson_2009G 20.41 29.94 <NA> <unk> wall across  
the entire african continent about the size of the great wall of china this  
would hardly be an invisible structure and yet it 's made  
MagnusLarsson_2009G 1 MagnusLarsson_2009G 29.9 31.68 <NA> <unk> from parts that  
are invisible <unk>  
MagnusLarsson_2009G 1 MagnusLarsson_2009G 32.71 36.93 <NA> to the naked eye  
bacteria and grains of sand  
MagnusLarsson_2009G 1 MagnusLarsson_2009G 37.07 46.3 <NA> now as architects we  
're trained to solve problems but i don 't really believe in architectural  
problems i only believe in opportunities which is why i 'll show you a threat  
MagnusLarsson_2009G 1 MagnusLarsson_2009G 46.3 48.69 <NA> and an architectural  
response <unk>
```

Abbildung 5.5: Beginn eines TED Transkripts

Zuerst wurden die Transkripte mit dem `TranscriptExtractor.py`¹⁸ Script extrahiert, Markierungen mit Bezug auf den Audiofile, wie sie in Abbildung 5.5 zu sehen sind, wurden entfernt und der reine Text wurde pro Talk zusammengefügt und als TXT-Datei abgespeichert. Anschließend wurde mit den in Unterunterabschnitt 5.1.2.1 beschriebenen Verfahren versucht, Interpunktions- in die Texte einzufügen. Auch hier führte kein Verfahren zu einem befriedigenden

¹⁸`TED/TranscriptExtractor.py`

Ergebnis, was in Tabelle 5.2 erkennbar ist. Obwohl der Versuch, Interpunktions einzufügen, scheiterte, war die Qualität der Transkripte dennoch gut genug, um sie für den Teil der Analyse zu nutzen, der sich nicht auf die Satzstruktur bezieht. So kann der TED-LIUM 3 Corpus wertvolle Informationen darüber liefern, welche und wie viele Diskursmarker in Reden verwendet werden und ob sie eher am Anfang oder am Ende der Rede vorkommen.

| Manuelle Interpunktions | | |
|--|---|--|
| Pausenbasiert | DeepSegment | BERT |
| over one point five billion people experience armed conflict. in response people are forced to flee their country leaving over fifteen million refugees. children without a doubt are the most innocent and vulnerable victims <unk> but not just from the obvious physical dangers <unk> but from the often unspoken effects that wars have on their families. <unk> the experiences of war leave children at a real high risk for the development of emotional and behavioral problems. children as we can only imagine will feel worried threatened and at risk but there is good news. | over one point five billion people experience armed conflict in response people are forced to flee their country leaving over fifteen million refugees children without a doubt are the most innocent and vulnerable victims <unk> but not just from the obvious physical dangers <unk> but from the often unspoken effects that wars have on their families <unk> the experiences of war leave children at a real high risk for the development of emotional and behavioral problems children as we can only imagine will feel worried threatened and at risk but there is good news | . over one point five billion people experience armed conflict and in response people are forced to flee their country leaving over fifteen million refugees . children without a doubt are the most innocent and vulnerable victims. many but not just from the obvious physical dangers, but from the often unspoken effects that wars have on their families. and the experiences of war leave children at a real high risk for the development of emotional and behavioral problems . children as we can only imagine will feel worried threatened and at risk but there is good news about |

Tabelle 5.2: Veranschaulichung der niedrigen WER des TED-LIUM 3 Corpus und der dennoch unzureichenden Qualität der eingefügten Interpunktions. Rot markiert sind die vom jeweiligen Verfahren eingefügten Symbole. Für BERT wurden hier nicht nur die Pausen, sondern auch die <unk> Tags maskiert und entsprechend vorhergesagt.

Da der Corpus keine weiteren Metadaten zu Genres oder Themen der einzelnen Reden enthält wurde entschieden, die Transkripte zusammenfassend unter dem Genre PRESENTATION und der Konversationsart REDE einzuordnen.

5.1.4 Zusammenfassung

Letztendlich wurden zwei Corpora gefunden, die ausreichend gute Daten liefern, um Diskursmarker in Texten zu analysieren, die für das akustische Medium konzipiert wurden: Der Spotify Podcast Corpus [Clifton et al., 2020] und der TED-LIUM 3 Corpus [Hernandez et al., 2018]. Beide Corpora fallen aufgrund ihrer Konzeption für das akustische Medium in die ORAL-AKUSTISCHE Kategorie der Diskursarten und können außerdem für weitergehende Analysen in verschiedene Genres und Konversationsarten eingeteilt werden. Der RadioTalk Corpus [Beeferman et al., 2019] wurde nach ausführlicher Betrachtung aufgrund von fehlender Interpunktions- und zu ungenauer Transkripte von der weiteren Verwendung ausgeschlossen. Eine Gegenüberstellung der Corpora im Bezug auf die in Unterabschnitt 5.1.1 aufgestellten Kriterien ist in Tabelle 5.3 zu sehen.

| Kategorie | TED-LIUM3 | Spotify | RadioTalk |
|-----------------------------|-----------|---------|-----------|
| Kostenlos | ✓ | ✓ | ✓ |
| Ausreichend relevante Daten | ✓ | ✓ | ✓ |
| Ausreichende Qualität | ✓ | ✓ | ✗ |
| Verschiedene Quellen | ✓ | ✓ | ✓ |
| Enthält Nachrichten | ✗ | ✓ | ? |
| Enthält Meinungen | ✓ | ✓ | ✓ |
| Manuell erstellt | ✗ | ✗ | ✗ |
| Genre-Informationen | ✗ | ✓ | ✗ |
| RSS-Feed | ✗ | ✓ | ✗ |
| Interpunktions- | ✗ | ✓ | ✗ |
| Word-Error-Rate | 6,7% | 18,1% | 13,1% |

Tabelle 5.3: Vergleich der akustischen Corpora

Tabelle 5.4 und Tabelle 5.5 zeigen die Überschneidungen der verschiedenen Genres und Konversationsarten in Bezug auf Episodenanzahl und Wortanzahl.

| Genres | Konversationsarten | | | |
|-----------------------|--------------------|---------|--------------|--------|
| | Dialog | Monolog | Coop-Monolog | Speech |
| News | 35 | 62 | 0 | 0 |
| Discussion | 216 | 4 | 0 | 0 |
| Science/ Education | 38 | 209 | 45 | 0 |
| Documentary | 38 | 44 | 2 091 | 0 |
| Presentation | 0 | 0 | 0 | 1 983 |

Tabelle 5.4: Heatmap der Überschneidungen von Episoden der einzelnen Genres mit den Konversationsarten

| Genres | Konversationsarten | | | |
|-----------------------|--------------------|-----------|--------------|-----------|
| | Dialog | Monolog | Coop-Monolog | Speech |
| News | 211 220 | 247 980 | 0 | 0 |
| Discussion | 2 123 195 | 24 944 | 0 | 0 |
| Science/ Education | 165 162 | 1 006 586 | 292 296 | 0 |
| Documentary | 462 431 | 124 935 | 12 605 933 | 0 |
| Presentation | 0 | 0 | 0 | 4 122 738 |

Tabelle 5.5: Heatmap der Wortanzahl in den gemeinsamen Episoden der einzelnen Genres mit den Konversationsarten

5.2 Schriftliche Corpora

5.2.1 Anforderungen

Um Diskursmarker in LITERAT-SCHRIFTLICHEM Diskurs zu untersuchen, wurde eine für den schriftlichen Diskurs konzipierte Datengrundlage benötigt. Auch diese Daten sollten möglichst aus dem Nachrichtenbereich stammen, was wesentlich einfacher zu realisieren war als bei den akustischen Corpora. Schriftliche Nachrichten-Corpora sind einfacher zusammenzustellen als Corpora mit Transkripten von Audiomaterial, da sie bereits in schriftlicher Form vorliegen. Sie

enthalten Interpunktions sowie Groß- und Kleinschreibung und haben naturgemäß eine WER von 0%. Zwei umfangreiche Nachrichten-Corpora sind der Gigaword Corpus [Parker et al., 2011] und der New York Times Corpus [Sandhaus, 2008]. Beide werden in Unterabschnitt 5.2.2 näher beschrieben.

5.2.2 Corpora

5.2.2.1 Gigaword Corpus

Der Gigaword Corpus [Parker et al., 2011] enthält Newswire Textdaten aus mehreren Jahren, bis zum Dezember 2010. Der Inhalt setzt sich zusammen aus sieben Quellen, die mit Wortanzahl und Dateigröße in Tabelle 5.6 angegeben sind. Insgesamt enthält der Corpus ca. vier Milliarden Wörter und fällt aufgrund seiner Konzeption für das schriftliche Medium in die LITERAT-SCHRIFTLICHE Kategorie der Diskursarten.

| Quelle | Wörter (tsd.) | MB |
|--|------------------|---------------|
| Agence France-Presse, English Service | 738 322 | 4 937 |
| Associated Press Worldstream, English Service | 1 186 955 | 7 889 |
| Central News Agency of Taiwan, English Service | 38 491 | 261 |
| Los Angeles Times/ Washington Post Newswire Service | 268 088 | 1 694 |
| Washington Post/ Bloomberg Newswire Service | 17 462 | 111 |
| New York Times Newswire Service | 1 422 670 | 8 938 |
| Xinhua News Agency, English Service | 360 714 | 2 518 |
| TOTAL | 4 032 686 | 26 348 |

Tabelle 5.6: Zusammensetzung und Größeneinschätzung des Gigaword Corpus

5.2.2.2 New York Times Corpus

Der New York Times Corpus [Sandhaus, 2008] ist ein annotierter Corpus. Er enthält über 1.8 Millionen Nachrichtenartikel, die zwischen dem 1. Januar 1987 und dem 19. Juni 2007 von der New York Times veröffentlicht wurden. Insgesamt beläuft sich der Corpus auf rund 1.1 Milliarden Wörter. Da alle Artikel für die schriftliche Ausgabe der New York Times verfasst wurden, eignet sich der New York Times Corpus bestens, um Diskursmarker im Hinblick auf ihr Vorkommen in für das schriftliche Medium konzipierten Texten zu analysieren und fällt in die LITERAT-SCHRIFTLICHE Kategorie der Diskursarten.

5.2.3 Zusammenfassung

Beide in Unterabschnitt 5.2.2 beschriebenen Corpora sind von guter Qualität und stellen eine ausreichende Menge an Nachrichtentexten zur Verfügung, die für das schriftliche Medium konzipiert wurden. Beide werden also in dieser Arbeit unter der LITERAT-SCHRIFTLICHEN Kategorie für die Analyse verwendet.

6

Feature Engineering

Die in diesem Kapitel vorgestellte Erweiterung einer bereits existierenden UIMA Pipeline¹ ermöglicht das Erkennen und Annotieren von Diskursmarkern in Texten. Auf den groben Aufbau der Pipeline und den Annotations-Prozess wird in Abschnitt 6.1 eingegangen. Abschnitt 6.2 beschreibt, wie Diskursmarker in den Texten erkannt werden und Abschnitt 6.3 geht auf die anhand der Annotationen berechneten Scores und deren Export für die Analyse ein.

Die Pfade zu den referenzierten Dateien sind relative Dateipfade im oben genannten Repository der Pipeline.

6.1 Aufbau der Analysis Pipeline

Das Ziel der Pipeline ist, in einer großen Menge unstrukturierter Texte Diskursmarker zu erkennen, im Text zu annotieren und basierend auf diesen Annotationen Scores für die Verteilung der DM im Text zu berechnen.

Die ursprüngliche Pipeline wurde als UIMA Framework konzipiert, damit der Code in unterschiedlichen Programmen und Analysen wiederverwendbar bleibt. Dadurch konnte sie ohne Probleme für diese Arbeit verwendet werden und enthielt bereits Module, die verschiedene Corpora einlesen und verarbeiten können. Für den Gigaword und den New York Times Corpus sind eigens erstellte Corpus Reader vorhanden² und da die Transkripte des Spotify und des TED-LIUM 3 Corpus, wie in den Unterabschnitten 5.1.3.2 und 5.1.3.3 beschrieben,

¹<https://git.webis.de/code-research/conversational-search/listenability-tools>, accessed 23.12.2020

²src/main/java/de/webis/corpus/reader

als reine Textdateien gespeichert wurden, konnten sie mit dem ebenfalls vorhandenen `PlainTextReader.java`³ eingelesen und verarbeitet werden.

In einem ersten Schritt wurde ein XML Parser für die Liste der Diskursmarker erstellt. Die Klasse `DiscourseMarkerXMLDOMReader.java`⁴ verarbeitet die EnDimLex XML-Datei und erstellt eine Map mit den Diskursmarkern als Keys und den jeweiligen orthografischen Variationen dieses DMs als Value-Liste. Als nächstes wurde der `DiscourseMarker.java`⁵ AnnotationType erstellt, der es ermöglicht, im Text Annotationen für DM hinzuzufügen.

Die Pipeline liest zunächst die Corpus-Texte ein und führt anschließend den `DiscourseMarkerAnalyser.java`⁶ aus. Hier werden die Diskursmarker eingelesen und mit den Texten gematched. Der Matching-Prozess wurde mit Hilfe des KMP String-Matching Algorithmus realisiert (siehe Abschnitt 6.2). Anhand der gefundenen Matches werden die Diskursmarker-Annotationen erstellt. Sie basieren auf der Position des Anfangs- und des Endzeichens des jeweiligen Matches im Text. Als nächstes wird die `DiscourseMarkerAnalysisEngine.java`⁷ ausgeführt. Hier werden aus den zuvor erstellten Annotationen, die angeben, an welcher Stelle im Text ein DM steht, die in Unterabschnitt 6.3.1 beschriebenen Scores für die Analyse berechnet und entsprechend als Annotationen zum Satz oder Text hinzugefügt. Alle Annotationen werden pro Dokument in einer XMI-Datei gespeichert.

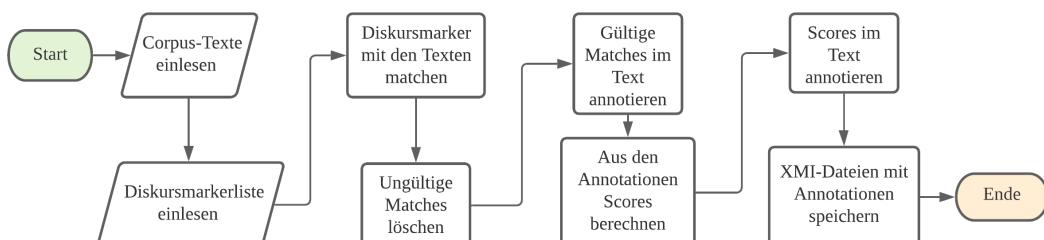


Abbildung 6.1: Programmablaufplan der Analysis Pipeline

³src/main/java/de/aitools/commons/uima/standard/PlainTextCollectionReader.java

⁴src/main/java/de/webis/discoursemarkers/DiscourseMarkerXMLDOMReader.java

⁵src/main/java/de/webis/writing/types/DiscourseMarker.java

⁶src/main/java/de/webis/listenability/feature/discoursemarker/DiscourseMarkerAnalyser.java

⁷src/main/java/de/webis/listenability/feature/discoursemarker/DiscourseMarkerAnalysisEngine.java

6.2 Diskursmarker erkennen durch String-Matching

Um Diskursmarker im Text zu erkennen, wurde ein einfaches String-Matching Verfahren gewählt. Der Knuth-Morris-Pratt String-Matching Algorithmus ist ein Verfahren mit linearer Laufzeit $\mathcal{O}(n+m)$, wobei n die Länge des Textes und m die Länge des gesuchten Musters ist. Da die Länge der EnDimLex Liste sich auf nur 149 Einträge beläuft, ist eine lineare Laufzeit auch bei einer großen Menge an Texten ausreichend gut.

Bei der Brute-Force Version eines String-Matching Algorithmus wird jede mögliche Position des gesuchten Musters im Text getestet. Das Muster wird Zeichen für Zeichen mit dem Text verglichen und bei einem Mismatch wird es um eine Stelle nach vorne verschoben, bevor der Matching-Prozess von neuem beginnt. Dies führt im Worst Case zu einer Zeitkomplexität von $\mathcal{O}(nm)$.

Der KMP-Algorithmus hingegen kann das Muster bei einem Mismatch auch um mehr als eine Stelle nach vorne verschieben, da bereits verglichene Zeichen in den Prozess mit einbezogen werden. Zuerst wird hierfür eine Präfix Tabelle erstellt, die das längste Präfix des Musters anzeigt, das gleichzeitig auch ein Suffix ist. Tabelle 6.1 zeigt eine beispielhafte Präfix Tabelle des Musters »onions«⁸. Das Matching beginnt wie bei der Brute-Force Version, doch wird bei jedem Mismatch in der Präfix Tabelle nachgesehen, um wie viele Stellen das Muster nach vorne verschoben werden kann, ohne ein Match zu überspringen. Dazu wird die Differenz des Index des Matches mit dem Wert des Index vor dem Mismatch berechnet und das Muster anschließend um entsprechend viele Stellen verschoben. Tabelle 6.2 zeigt ein beispielhaftes Mismatch in einem Matching-Prozess. Das Mismatch taucht hier an Index 5 des Musters auf, also wird in Tabelle 6.1 nachgesehen, welchen Wert Index 4 hat. Der Wert ist in diesem Fall 2 und das Muster wird demnach um $5 - 2 = 3$ Stellen nach vorne verschoben, was überflüssige Vergleiche ausschließt.

Auf diese Weise kann der KMP-Algorithmus wesentlich schneller sein als die Brute-Force Variante, wenn nach Mustern gesucht wird, die Präfixe enthalten, welche gleichzeitig auch Suffixe sind. Gegenüber anderen Verfahren wie bei-

⁸<https://www.youtube.com/watch?v=4jY57Ehc14Y&t=640s>, accessed 31.12.2020

| Index | 0 | 1 | 2 | 3 | 4 | 5 |
|-------------|---|---|---|---|---|---|
| Muster | o | n | i | o | n | s |
| Präfixlänge | 0 | 0 | 0 | 1 | 2 | 0 |

Tabelle 6.1: Beispiel einer Präfix Tabelle für den KMP-Algorithmus

| Match | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | | | |
|--------|---|---|---|---|---|---|---|---|---|
| Text | o | n | i | o | n | i | o | n | s |
| Muster | o | n | i | o | n | s | | | |

Tabelle 6.2: Beispielhafter Schritt eines Matching-Prozesses

spielsweise dem Generalized Suffix Tree⁹ hat der KMP-Algorithmus außerdem den Vorteil, dass die Suche nach Phrasen, also nach mehreren zusammenhängenden Wörtern wie z.B. *on the other hand*, möglich ist.

6.2.1 Probleme

6.2.1.1 Substring Matches

Ein Problem des String-Matchings ist die Tatsache, dass auch Substrings von Wörtern gematched werden. Beispielsweise wird in dem Wort *reasoning* der DM *so* gematched, obwohl die Silbe *so* in diesem Fall natürlich kein DM und somit auch kein korrektes Match ist. Um dieses Problem zu lösen, wurde eine Treemap aller Wörter des jeweiligen Textes erstellt, die als Keys die jeweilige Anfangsposition und als Values die entsprechende Endposition des Wortes im Text enthält. Nach Vollendung des Matching-Prozesses wurde mit der `eliminateSubmatches()`¹⁰ Funktion für jedes Match überprüft, ob es mit Anfangs- und Endposition in der Treemap enthalten war. Falls nicht, konnte das entsprechende Match als nicht vollständiges Wort und damit inkorrekt Match gelöscht werden.

⁹https://en.wikipedia.org/wiki/Generalized_suffix_tree, accessed 31.12.2020

¹⁰`src/main/java/de/webis/discoursemarkers/DiscourseMarkerMatches.java`

6.2.1.2 Doppelte Matches bei DM Phrasen

Ein ähnliches Problem tritt bei Phrasen auf, die ein DM sind und gleichzeitig einen oder mehrere weitere DM enthalten. So gäbe es für die Phrase *rather than* beispielsweise zwei Matches: Einen für die gesamte Phrase und einen für den enthaltenen DM *rather*. Die Zählung beider Matches würde das Ergebnis verfälschen, da in diesem Fall *rather* kein eigener DM ist, sondern nur Teil der verwendeten DM-Phrase *rather than*. Daher wurde entschieden, immer die längste Phrase als Match zu betrachten und doppelte Matches für enthaltene DM oder kürzere enthaltene DM Phrasen herauszufiltern. Dazu wurde nach abgeschlossener Suche jedes Match mit der `eliminateDoubles()`¹¹ Funktion anhand der Anfangs- und Endposition der Matches daraufhin untersucht, ob es ein weiteres Match enthält. War dies der Fall, wurde dieses enthaltene Match gelöscht.

6.3 Scores

Die Analysis Pipeline nutzt die erstellten Diskursmarker-Annotationen, um Scores für die Verteilung der DM zu generieren. Diese Scores haben einen *Namen*, einen *Value* und eine optionale *Explanation*. Der Name ist ein String und kann frei definiert werden, er dient der Unterscheidung. Der Value ist ein Zahlenwert, der auf Grundlage der Positionierung der jeweiligen Annotation im Text berechnet wird. Beispielsweise kann die Position eines DMs im Satz berechnet werden, indem das erste Zeichen des Matches relativ zum ersten Zeichen des enthaltenden Satzes betrachtet wird. Die Explanation dient der weiteren Ausführung oder Erklärung eines Scores. Auch sie ist ein String und kann beliebig definiert werden.

6.3.1 Annotierte Scores

Jeder Satz eines Dokuments wird mit den in Tabelle 6.3 beschriebenen Scores und zugehöriger Explanation annotiert. Jedes Dokument erhält außerdem die in Tabelle 6.4 beschriebenen Score-Annotationen.

¹¹ `src/main/java/de/webis/discoursemarkers/DiscourseMarkerMatches.java`

| Score | Funktion |
|------------------|--|
| dm_count | Anzahl der DM im Satz <i>Explanation:</i> Enthaltene DM |
| dm_pos_sent_perc | Position jedes DMs im Satz in % <i>Explanation:</i> Enthaltene DM |

Tabelle 6.3: Annotierte Scores für jeden Satz eines Dokuments

| Score | Funktion |
|--------------------|---|
| word_count_doc | Anzahl der Wörter im Dokument |
| dm_count_doc | Anzahl der DM im Dokument |
| sentence_count_doc | Anzahl der Sätze im Dokument |
| dm_sentences | Anzahl der Sätze im Dokument, die mind. einen DM enthalten |
| dm_pos_doc_perc | Positionen der DM im Dokument in % <i>Explanation:</i> Die entsprechenden DM |

Tabelle 6.4: Annotierte Scores für jedes Dokument

6.3.2 Exportierte Werte

Von der Analysis Pipeline wird pro Dokument eine XMI-Datei mit allen Annotationen erstellt. Das separate Programm `ScoresToCsv.java`¹² kann diese Dateien einlesen und alle zur weiteren Analyse benötigten Werte berechnen und exportieren. Für jeden Corpus wurde mit diesem Programm eine CSV-Tabelle



Abbildung 6.2: Programmablaufplan der `ScoresToCsv.java`

erstellt, die pro Dokument eine Zeile mit den entsprechend für das Dokument berechneten Werten enthält. Die aus den annotierten Scores berechneten Werte sind in Tabelle 6.5 beschrieben.

¹²<src/main/java/de/webis/listenability/ScoresToCsv.java>

| Wert | Funktion |
|--------------------|---|
| document | Dateiname des Dokuments |
| word_count_doc | Anzahl der Wörter |
| dm_count_doc | Anzahl der DM |
| dm_words_perc | %-Anteil der DM an allen Wörtern des Dokuments |
| sentence_count_doc | Anzahl der Sätze |
| dm_sentences | Anzahl der Sätze, die mind. einen DM enthalten |
| dm_sentences_perc | %-Anteil der Sätze, die mind. einen DM enthalten, an allen Sätzen des Dokuments |
| dm_pos_sent_begin | Anzahl der DM, die am Satzanfang stehen |
| dm_pos_sent_middle | Anzahl der DM, die in der Satzmitte stehen |
| dm_pos_sent_end | Anzahl der DM, die am Satzende stehen |
| dm_pos_doc_begin | Anzahl der DM, die am Dokumentanfang stehen |
| dm_pos_doc_middle | Anzahl der DM, die in der Dokumentmitte stehen |
| dm_pos_doc_end | Anzahl der DM, die am Dokumentende stehen |
| dm_count_sent | String in Form eines Python-Dictionarys, das als Key Anzahlen von DM enthält und als jeweiligen Value die Anzahl an Sätzen, die entsprechend viele DM enthalten. Ein Key könnte beispielsweise 3 sein und der Value gibt dann an, wie viele der Sätze im Dokument genau 3 DM enthalten. |
| dm_counts_dict | String in Form eines Python-Dictionarys, das als Keys die im Dokument enthaltenen DM enthält und als Values je die folgenden Werte: |

Fortsetzung auf nächster Seite ...

| Wert | Funktion |
|------|--|
| | total Gesamtes Vorkommen im Dokument |
| | sent_begin Anzahl der Vorkommen am Satzanfang |
| | sent_middle Anzahl der Vorkommen in der Satzmitte |
| | sent_end Anzahl der Vorkommen am Satzende |
| | doc_begin Anzahl der Vorkommen am Dokumentanfang |
| | doc_middle Anzahl der Vorkommen in der Dokumentmitte |
| | doc_end Anzahl der Vorkommen am Dokumentende |

Tabelle 6.5: Exportierte Scores und ihre Bedeutung

6.3.2.1 Einteilung der Sätze und Texte in Anfang, Mitte und Ende

Tabelle 6.5 zeigt Werte für die Positionen Anfang, Mitte und Ende im Satz oder Dokument. Diese Werte wurden unter der Annahme berechnet, dass die ersten 15% der Zeichen eines Satzes oder Dokumentes jeweils den Anfang darstellen und die letzten 15% das Ende. Die restlichen 70% der Zeichen werden als Mitte des Satzes oder Dokuments betrachtet [Syed et al., 2020].

Diese Annahme soll mit Beispielen¹⁸ zur Satzposition illustriert werden:

(1) *Nevertheless to Isengard I go.*

Dieser Satz hat 30 Zeichen (inkl. Interpunktions- und Leerzeichen). 15% sind entsprechend alle Zeichen bis einschließlich des fünften. Das erste Zeichen des DMs *nevertheless* ist gleichzeitig das erste Zeichen des Satzes, also Teil der ersten 15%, folglich steht der DM am Anfang des Satzes. Dies entspricht der intuitiven Annahme über die Position von *nevertheless*.

In folgendem Beispiel ist die intuitive Annahme, dass *however* am Satzanfang positioniert ist, *and* hingegen in der Mitte:

(2) *Gandalf, however, disbelieved Bilbo's first story, as soon as he heard it, and he continued to be very curious about the ring.*

¹⁸Beispielsätze entnommen aus *The Lord of the Rings* von J.R.R. Tolkien, veröffentlicht von HarperCollinsPublishers, 2005

Dieser Satz besteht aus 126 Zeichen, 15% davon sind alle Zeichen bis einschließlich Zeichen 19. Das Anfangszeichen von *however* befindet sich an Position 10 des Satzes¹⁴, folglich am Satzanfang. Die Mitte des Satzes sind die nächsten 70% der Zeichen, also die Positionen 20 bis einschließlich 89. Das Anfangszeichen von *and* befindet sich an Position 76 des Satzes, folglich in der Mitte des Satzes. Beide Positionierungen entsprechen den zuvor getroffenen Annahmen.

Gleiches gilt für das Satzende. Der DM *anyway* wird im folgenden Beispiel intuitiv am Satzende angenommen:

(3) I wonder what Ents can do about it *anyway*.

Da dieser Satz aus 42 Zeichen besteht, beginnen die letzten 15%, das Satzende, an Position 36. Dies ist exakt die Position, an der auch *anyway* beginnt. Damit wird *anyway* korrekt am Satzende positioniert.

Diese Art der Einteilung wurde an einer Reihe weiterer Beispiele getestet und die Ergebnisse deuten darauf hin, dass sie für die hier durchgeführte empirische Untersuchung ausreichend genau ist.

¹⁴Zählung beginnend von 1, da das erste Zeichen des Satzes/Wortes gemeint ist

Auswertung

Die durch das in Kapitel 6 beschriebene Verfahren gewonnenen Scores wurden mithilfe von Python Scripten¹ statistisch ausgewertet und visualisiert. Im Folgenden werden die anfangs aufgestellten Forschungsfragen unter Bezugnahme auf die so erstellten Diagramme und Statistiken betrachtet. Für jede der fünf Fragen werden drei Vergleiche angestellt: Einer zwischen den Diskursarten, einer zwischen den Genres und einer zwischen den Konversationsarten. Die gezeigten Diagramme dienen der visuellen Unterstützung, während die statistische Auswertung sich auf die in Anhang B zu findenden Tabellen bezieht.

Statistik

Für die folgenden Betrachtungen wurde die Nullhypothese H_0 aufgestellt, dass im Bezug auf die entsprechende Frage kein Unterschied zwischen den jeweiligen Textsorten besteht. Um H_0 zu bestätigen oder zu verwerfen, werden die statistischen Größen *P-Wert* und *Effektgröße* (EG, *Cohens D*²) benutzt. Der P-Wert gibt an, ob die Ergebnisse für oder gegen H_0 sprechen: ein kleiner P-Wert suggeriert, dass die Ergebnisse H_0 nicht unterstützen. Ist der P-Wert kleiner als ein zuvor festgelegtes Signifikanzniveau α , ist das Testergebnis *statistisch signifikant*. In dieser Betrachtung wird der P-Wert gegen $\alpha = 0.05$ getestet. Der P-Wert wurde mit Welch's T-Test berechnet³. Alle P-Werte wurden außerdem Bonferroni-korrigiert. Der *Mean* gibt das arithmetische Mittel an.

¹<https://git.webis.de/code-teaching/theses/thesis-sacher/-/tree/master/scripts/analysis>

²https://matheguru.com/stochastik/effektstarke.html#Cohens_d-1, accessed 1.1.2021

³https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html, accessed 3.1.2021

Anmerkung: Generell ist zu beachten, dass der TED-LIUM 3 Datensatz, der in die Diskursart ORAL-AKUSTISCH fällt und außerdem die Textsorten PRESENTATION und REDE repräsentiert, von allen Fragestellungen, die sich auf die Satzstruktur beziehen, ausgeschlossen wurde, da dieser Datensatz keine Interpunktionsenthalt.

7.1 Welche Textsorten stützen sich besonders auf Diskursmarker?

Diese Frage betrachtet, welche Textsorten generell mehr Diskursmarker verwenden als andere. Dabei wird sowohl ein Blick auf die Menge der genutzten DM pro Text als auch pro Satz geworfen.

Die Ergebnisse der statistischen Tests für diesen Abschnitt sind in den Tabellen B.1 und B.2 zu finden.

7.1.1 Vergleich zwischen Diskursarten

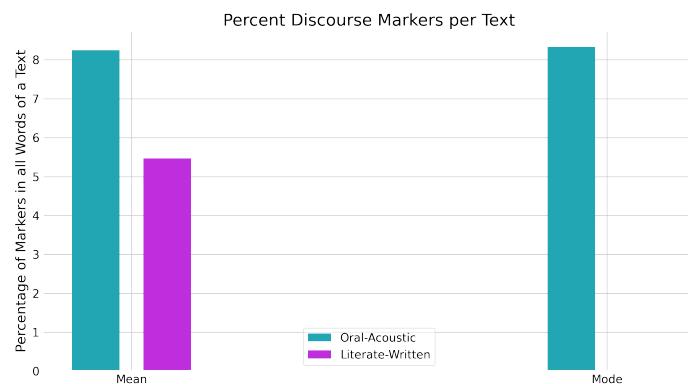


Abbildung 7.1: Prozentualer Anteil der DM an allen Wörtern der Texte

Anhand der Abbildungen 7.1 und 7.2 lässt sich gut erkennen, dass ORAL-AKUSTISCHE (OA) Texte im Durchschnitt mehr DM pro Text enthalten als LITERAT-SCHRIFTLICHE (LS) Texte (8% vs. 5% aller Wörter). Dieser Unterschied ist mit einem P-Wert < 0.001 statistisch signifikant. Die Effektgröße (EG) beträgt 1.6.

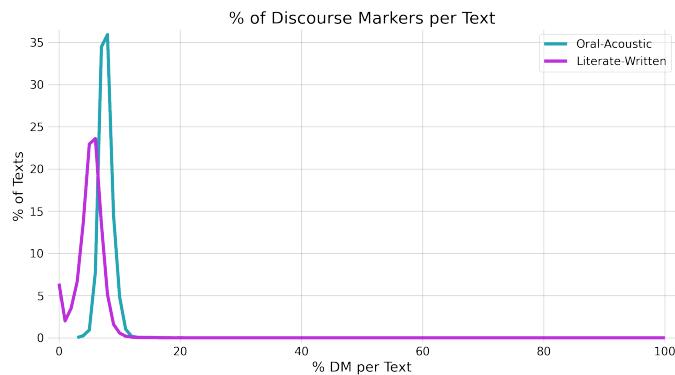


Abbildung 7.2: Vergleich des prozentualen Anteils der DM an allen Wörtern der Texte

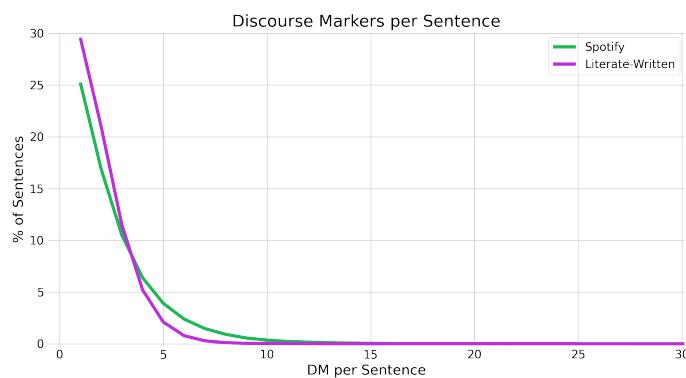


Abbildung 7.3: Vergleich der DM Anzahl pro Satz

Auch bei der DM Anzahl pro Satz lässt sich ein Unterschied erkennen. Der Spotify-Corpus, repräsentativ für die oral-akustischen Texte, enthält durchschnittlich 2.67 DM pro Satz, während es bei den literat-schriftlichen Texten 2.06 DM pro Satz sind. Wie in Abbildung 7.3 zu sehen, fällt die Kurve der pro Satz enthaltenen DM für den Spotify Corpus flacher aus als für die LS Texte. Auch dieser Unterschied ist mit einem P-Wert < 0.001 statistisch signifikant. Es wird deutlich, dass OA Texte signifikant mehr DM enthalten als LS Texte.

ORAL-AKUSTISCHE Texte stützen sich folglich mehr auf die Nutzung von Diskursmarkern als LITERAT-SCHRIFTLICHE Texte.

7.1.2 Vergleich zwischen Genres

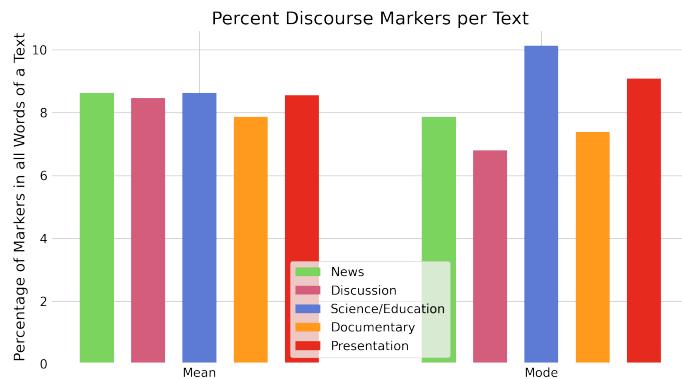


Abbildung 7.4: Prozentualer Anteil der DM an allen Wörtern der Texte

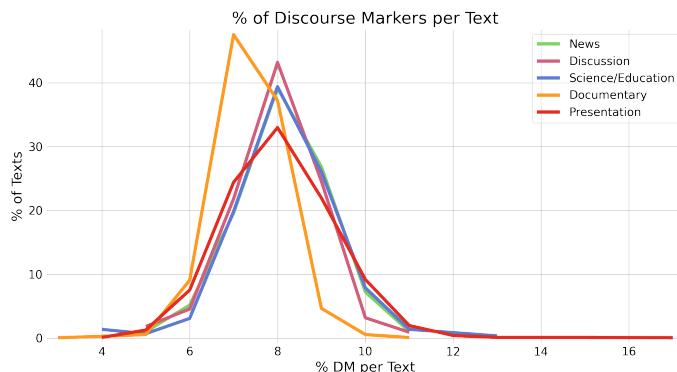


Abbildung 7.5: Vergleich des prozentualen Anteils der DM an allen Wörtern der Texte

Abbildung 7.4 gibt einen guten Überblick über die Verteilung der DM innerhalb der betrachteten Genres. Es ist festzustellen, dass das DOCUMENTARY Genre signifikant weniger DM enthält als alle anderen Genres (s. Abbildung 7.5), der P-Wert ist in allen Fällen < 0.001 . DOCUMENTARIES sind, wie in Kapitel 4 beschrieben, gut recherchierte, geskriptete Sendungen, die außer gelegentlichen Zeugenaussagen oder Interviews keine freie Rede enthalten. Sie sind bewusst für das akustische Medium konzipiert, während NEWS, DISCUSSION und SCIENCE/EDUCATION zum Großteil improvisiert sind. Auch diese Genres sind natürlich ORALAKUSTISCH, doch sind sie dies wegen Improvisation und freier Rede nicht durch

ein intentionales Konzept, sondern aufgrund ihrer Natur. Es fällt auf, dass sich das PRESENTATION Genre, das ausschließlich aus (geskripteten) TED-Talks besteht, nicht auf eine ähnliche Weise von den übrigen Genres abhebt. Dieser Umstand lässt sich möglicherweise damit erklären, dass innerhalb dieser Talks häufig vom Skript abgewichen und improvisiert wird oder die Skripte absichtlich so konzipiert sind, dass sie eher auf Umgangssprache setzen. Die von Schneider [2011] beschriebene Interaktion mit dem Publikum mag hier eine entscheidende Rolle spielen.

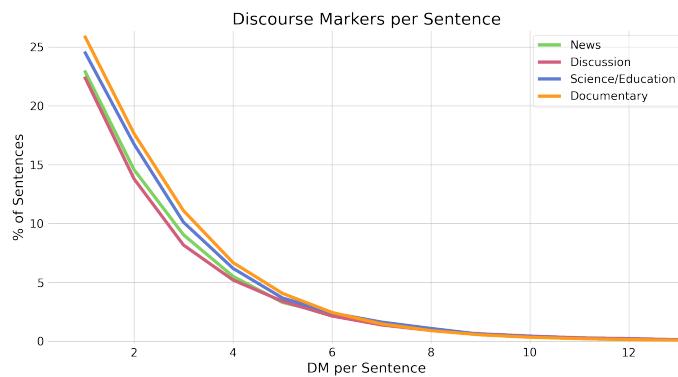


Abbildung 7.6: Vergleich der DM Anzahl pro Satz

Abbildung 7.6 zeigt die DM Anzahlen pro Satz. Auch hier hebt DOCUMENTARY sich mit signifikant weniger DM pro Satz deutlich von allen anderen Genres ab (P -Werte < 0.001).

Zusammenfassend wird deutlich, dass die Menge der genutzten DM nicht so sehr vom Genre abhängen scheint, sondern eher von der Art der Aufbereitung. DOCUMENTARIES, die strikt einem Skript folgen, nutzen signifikant weniger DM als Genres mit mehr improvisierten und freien Anteilen. In diesem Zusammenhang wäre es besonders hilfreich, die tatsächliche Listenability der DOCUMENTARIES messen und mit der der anderen Genres vergleichen zu können, doch war dies im Rahmen dieser Arbeit nicht mehr möglich.

7.1.3 Vergleich zwischen Konversationsarten

Bei den Konversationsarten treten zwischen fast allen Textsorten signifikante Unterschiede auf. Abbildung 7.7 gibt einen guten Überblick.

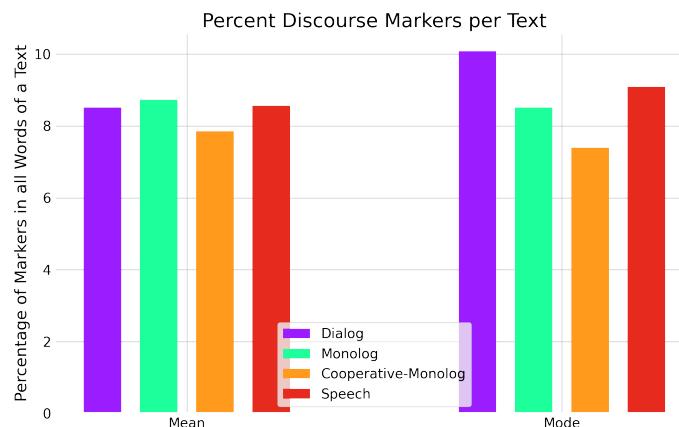


Abbildung 7.7: Prozentualer Anteil der DM an allen Wörtern der Texte

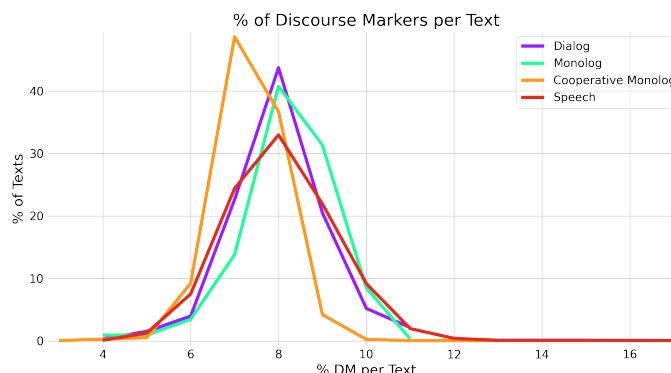


Abbildung 7.8: Vergleich des prozentualen Anteils der DM an allen Wörtern der Texte

Einzig zwischen DIALOG und REDE ist kein signifikanter Unterschied erkennbar. Die Ähnlichkeit zwischen diesen beiden Textsorten kann in der von Schneider [2011] angesprochenen Interaktion begründet liegen, die sowohl beim Dialog (Interaktion mit den Dialogpartnern) als auch bei der Rede (Interaktion mit dem Publikum) eine große Rolle spielt. Alle übrigen Konversationsarten weisen untereinander signifikante Unterschiede in der Menge der genutzten DM

auf. Hervorstechend ist der KOOPERATIVE MONOLOG, der signifikant weniger DM nutzt als alle anderen Konversationsarten, gut zu sehen in Abbildung 7.8. Dies überrascht nach der Genre-Betrachtung in Unterabschnitt 7.1.2 nicht, sind doch die Texte des KOOPERATIVEN MONOLOGS nahezu identisch mit denen des Genres DOCUMENTARY (vgl. auch die Tabellen 5.4 und 5.5).

Der MONOLOG hebt sich sowohl vom DIALOG als auch von der REDE ab und nutzt signifikant mehr DM als die beiden anderen Konversationsarten.

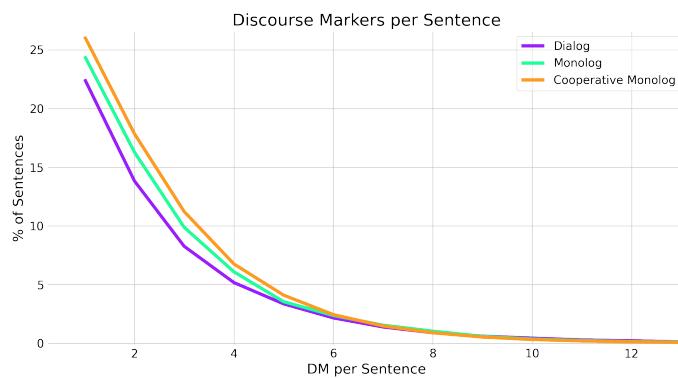


Abbildung 7.9: Vergleich der DM Anzahl pro Satz

Pro Satz jedoch werden im DIALOG signifikant mehr DM genutzt als im MONOLOG ($p < 0.001$, EG: 0.04) oder im KOOPERATIVEN MONOLOG ($p < 0.001$, EG: 0.06). Auch im MONOLOG werden pro Satz mehr DM genutzt als im KOOPERATIVEN MONOLOG, was letzteren auch bei der DM Nutzung pro Satz mit den wenigsten DM heraus stechen lässt.

Zusammenfassend wird deutlich, dass der geskriptete und geplante KOOPERATIVE MONOLOG signifikant weniger DM nutzt als alle anderen Konversationsarten. Außerdem hebt sich der MONOLOG mit signifikant mehr genutzten DM pro Text deutlich von REDE und DIALOG ab.

7.2 An welchen Positionen im Text stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?

Betrachtet werden die Textpositionen wie in Unterabschnitt 6.3.2.1 beschrieben.

Die Ergebnisse der statistischen Tests für diesen Abschnitt sind in Tabelle B.3 zu finden.

7.2.1 Vergleich zwischen Diskursarten

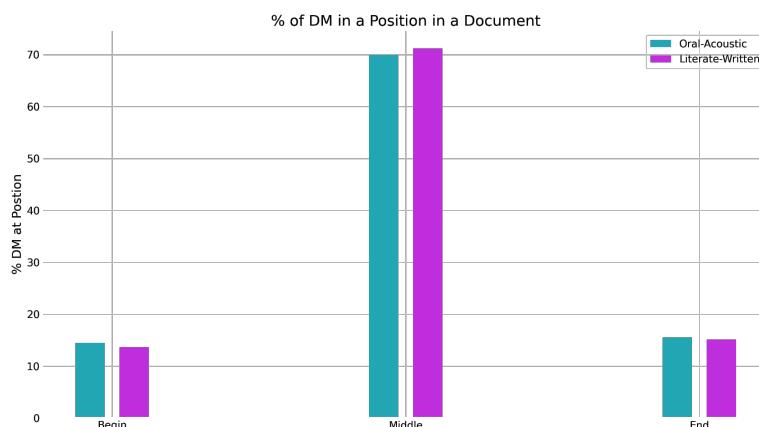


Abbildung 7.10: Prozentualer Anteil der DM an bestimmten Positionen im Text an allen DM des Textes

Abbildung 7.10 zeigt, dass die DM für beide Diskursarten recht gleichmäßig über den Text verteilt sind. Dennoch lässt sich für alle drei Textpositionen ein signifikanter Unterschied in der Menge der genutzten DM zwischen den Diskursarten feststellen. Am Textanfang sowie -ende enthalten OA Texte signifikant mehr DM als LS Texte (P -Werte < 0.001), während LS Texte in der Mitte signifikant mehr DM enthalten als OA Texte (p < 0.001).

ORAL-AKUSTISCHE Texte stützen sich folglich am Textanfang und am Textende mehr auf DM als LITERAT-SCHRIFTLICHE Texte, während diese DM eher in der Mitte einzusetzen als ORAL-AKUSTISCHE Texte.

7.2.2 Vergleich zwischen Genres

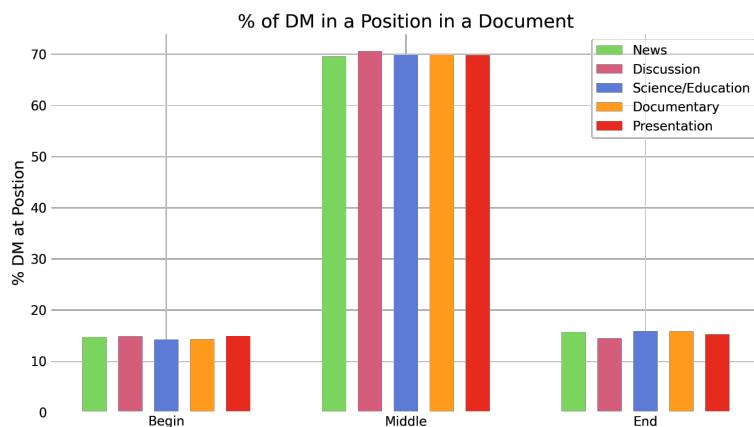


Abbildung 7.11: Prozentualer Anteil der DM an bestimmten Positionen im Text an allen DM des Textes

Abbildung 7.11 zeigt, dass die DM auch innerhalb der Genres jeweils sehr gleichmäßig über den gesamten Text verteilt sind.

Am Textanfang nutzen Texte des Genres DISCUSSION signifikant mehr DM als SCIENCE/EDUCATION ($p < 0.001$, EG: 0.3) oder DOCUMENTARY ($p < 0.001$, EG: 0.36). Auch PRESENTATION nutzt an dieser Stelle signifikant mehr DM als die beiden genannten Genres. Ansonsten lassen sich am Textanfang keine statistisch signifikanten Unterschiede zwischen den Genres feststellen.

In der Textmitte sind ähnliche Unterschiede zu beobachten: Wieder nutzt DISCUSSION an dieser Stelle mehr DM als SCIENCE/EDUCATION ($p < 0.001$, EG: 0.32) und DOCUMENTARY ($p < 0.001$, EG: 0.25). Weitere signifikante Unterschiede lassen sich an dieser Position nicht feststellen.

Bei den DM am Ende des Textes nutzen sowohl SCIENCE/EDUCATION als auch DOCUMENTARY jeweils signifikant mehr DM als DISCUSSION. Wie schon am Textanfang hebt sich hier wieder PRESENTATION ab und nutzt signifikant weniger DM als SCIENCE/EDUCATION und DOCUMENTARY.

Die interaktionslastigeren Genres DISCUSSION und PRESENTATION scheinen sich also vor allem am Textanfang und in der Mitte auf DM zu stützen. Am Ende des Textes sind es vor allem die sachlicheren Genres SCIENCE/EDUCATION und DOCUMENTARY, die von DM Gebrauch machen. PRESENTATION nutzt an dieser Stelle

signifikant weniger DM als andere Genres.

7.2.3 Vergleich zwischen Konversationsarten

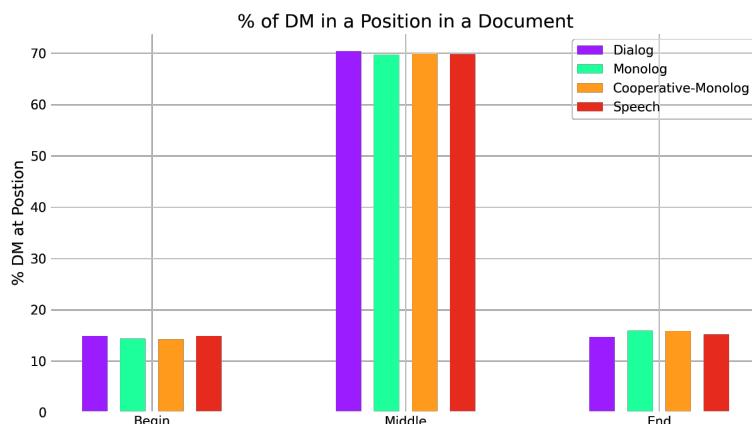


Abbildung 7.12: Prozentualer Anteil der DM an bestimmten Positionen im Text an allen DM des Textes

Auch bei den Konversationsarten sind die DM für alle Konversationsarten recht gleichmäßig über den gesamten Text verteilt.

Am Textanfang heben sich DIALOG und REDE hervor. Im DIALOG werden hier signifikant mehr DM genutzt als im MONOLOG (p: 0.006, EG: 0.22) und im KOOPERATIVEN MONOLOG (p: < 0.001, EG: 0.34). In der REDE hingegen werden signifikant weniger DM genutzt als im MONOLOG (p: 0.001, EG: 0.17) und im KOOPERATIVEN MONOLOG (p: < 0.001, EG: 0.27).

In der Textmitte sticht einzig der DIALOG hervor. Hier werden mehr DM genutzt als im MONOLOG (p: 0.006, EG: 0.3) und im KOOPERATIVEN MONOLOG (p: < 0.003, EG: 0.19).

Am Textende sind signifikante Unterschiede zwischen fast allen Konversationsarten feststellbar. Einzig MONOLOG und KOOPERATIVER MONOLOG unterscheiden sich an dieser Stelle nicht. Im DIALOG werden am Ende der Texte signifikant weniger DM genutzt als im MONOLOG (p < 0.001, EG: 0.56), KOOPERATIVEN MONOLOG (p < 0.001, EG: 0.58) und der REDE (p: 0.002, EG: 0.17). In der REDE wiederum werden signifikant weniger DM genutzt als im MONOLOG (p < 0.001, EG: 0.33) und im KOOPERATIVER MONOLOG (p < 0.001, EG: 0.31).

Die interaktiven Konversationsarten REDE und DIALOG stützen sich also vor allem am Textanfang auf DM, im Gegensatz zum MONOLOG und KOOPERATIVEN MONOLOG. Diese Konversationsarten hingegen setzen am Ende des Textes eher auf DM als die beiden erstgenannten.

7.3 An welchen Positionen im Satz stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?

Betrachtet werden die Satzpositionen wie in Unterabschnitt 6.3.2.1 beschrieben.

Da diese Fragestellung sich auf die Satzstruktur bezieht, wurde der TED-LIUM 3 Datensatz von der Analyse ausgeschlossen. Repräsentativ für die ORAL-AKUSTISCHE Diskursart wird der Spotify Datensatz verwendet. Das Genre PRESENTATION und die Konversationsart REDE fallen weg.

Die Ergebnisse der statistischen Tests für diesen Abschnitt sind in Tabelle B.4 zu finden.

7.3.1 Vergleich zwischen Diskursarten

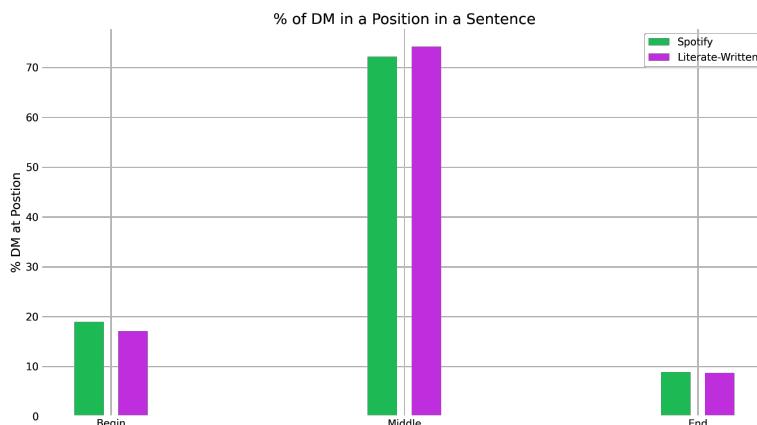


Abbildung 7.13: Prozentualer Anteil der DM an bestimmten Positionen im Satz an allen DM des Textes

Abbildung 7.13 zeigt, dass die DM für beide Diskursarten gleichmäßig über den ganzen Satz verteilt sind.

Am Satzanfang nutzen ORAL-AKUSTISCHE Texte signifikant mehr DM als LITERAT-SCHRIFTLICHE ($p < 0.001$, EG: 0.38), während LS Texte wiederum in der Satzmitte signifikant mehr DM nutzen als OA Texte ($p < 0.001$, EG: 0.34). Am Satzende lässt sich im Bezug auf die Menge der genutzten DM kein statistisch signifikanter Unterschied feststellen.

7.3.2 Vergleich zwischen Genres

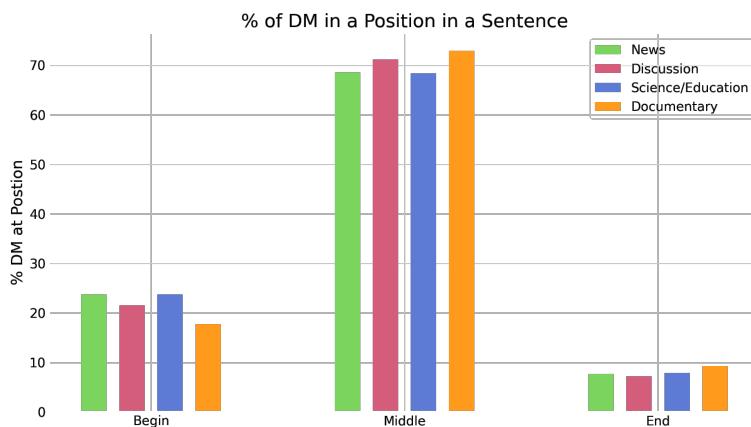


Abbildung 7.14: Prozentualer Anteil der DM an bestimmten Positionen im Satz an allen DM des Textes

Bei der Betrachtung von Abbildung 7.14 fällt auf, dass die DM für die Genres nicht ganz gleichmäßig über den Satz verteilt sind. Bei allen Genres werden am Satzanfang mehr DM genutzt als am Satzende.

Am Satzanfang besteht nur zwischen NEWS und SCIENCE/EDUCATION kein statistisch signifikanter Unterschied. Ansonsten enthalten NEWS an dieser Stelle mehr DM als DISCUSSION ($p < 0.001$, EG: 0.54) und als DOCUMENTARY ($p < 0.001$, EG: 1.47). Auch SCIENCE/EDUCATION enthält mehr DM als DISCUSSION ($p < 0.001$, EG: 0.48), während DOCUMENTARY weniger DM enthält als DISCUSSION ($p < 0.001$, EG: 1.23) und auch weniger als SCIENCE/EDUCATION ($p < 0.001$, EG: 1.34). Ein exakt umgekehrtes Bild im Bezug auf mehr/weniger DM bietet sich in der Satzmitte. Am Satzende werden in DOCUMENTARY signifikant weniger DM

verwendet als in allen anderen Genres.

Es fällt auf, dass die sachlicheren Genres NEWS und SCIENCE/EDUCATION sich eher am Satzanfang auf DM stützen als die übrigen Genres. Dies kehrt sich in der Satzmitte um. In allen Fällen stützen sich die geskripteten Texte des DOCUMENTARY-Genres weniger auf DM als die anderen Genres.

7.3.3 Vergleich zwischen Konversationsarten

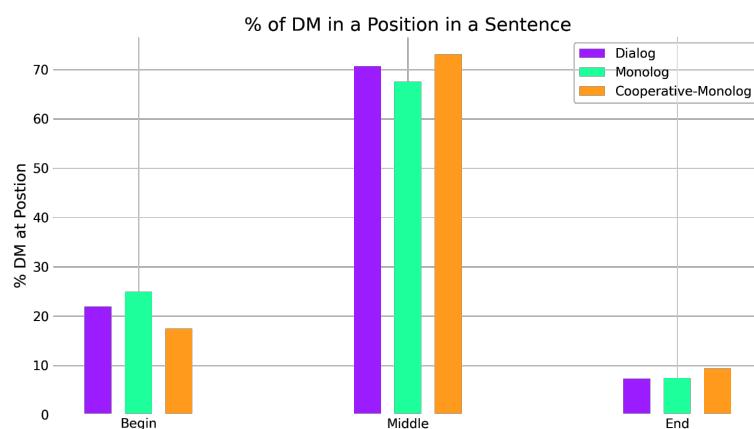


Abbildung 7.15: Prozentualer Anteil der DM an bestimmten Positionen im Satz an allen DM des Textes

Abbildung 7.15 zeigt, dass auch bei den Konversationsarten am Satzanfang wesentlich mehr DM verwendet werden als am Satzende.

Sowohl am Satzanfang als auch in der Satzmitte lassen sich signifikante Unterschiede in der Anzahl der verwendeten DM zwischen allen Konversationsarten feststellen. Am Satzanfang enthält der DIALOG weniger DM als der MONOLOG, dies kehrt sich in der Satzmitte um und am Satzende ist kein Unterschied erkennbar. Texte des KOOPERATIVEN MONOLOGS enthalten am Satzanfang weniger DM als DIALOG ($p < 0.001$, EG: 1.35) und MONOLOG ($p < 0.001$, EG: 1.62). Auch diese Verhältnisse kehren sich für die Satzmitte und das Satzende um und der KOOPERATIVE MONOLOG nutzt hier jeweils mehr DM als MONOLOG und DIALOG.

Am Satzanfang stützen sich also vor allem Texte des DIALOGS auf DM, während am Satzende Texte des KOOPERATIVEN MONOLOGS signifikant mehr DM nut-

zen als die anderen Konversationsarten.

7.4 Auf welche Klassen von Diskursmarkern stützen sich die jeweiligen Textsorten besonders?

Diese Fragestellung bezieht sich auf die in Abschnitt 3.2 definierten DM-Klassen TEMPORAL, CONTINGENCY, COMPARISON und EXPANSION. Abbildung 7.16 zeigt die Verteilung aller 149 DM auf diese Klassen. Die vollständige Tabelle der betrachteten DM kann in Tabelle A.1 eingesehen werden.

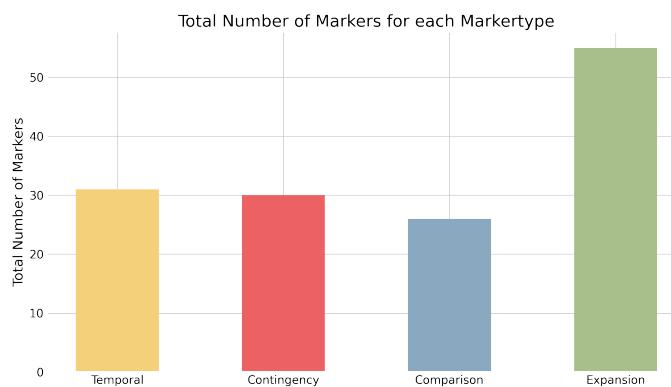


Abbildung 7.16: Aufteilung der 149 DM in Typen, vgl. auch Tabelle A.1

Die Ergebnisse der statistischen Tests für diesen Abschnitt sind in Tabelle B.5 zu finden.

7.4.1 Vergleich zwischen Diskursarten

Die Abbildungen 7.17 und 7.18 zeigen deutlich, dass EXPANSION DM in beiden Diskursarten den wesentlich größten Anteil aller DM ausmachen. Betrachtet man jedoch Abbildung 7.16, sieht man, dass diese Gruppe auch die meisten DM enthält, u.a. den DM *and*, was den Anteil von ca. 50% an allen DM bei beiden Diskursarten erklären kann (s. Abbildung 7.18b).

Abbildung 7.18 zeigt den prozentualen Anteil einer jeden DM-Klasse an allen DM der jeweiligen Diskursart. Für die Klassen TEMPORAL und EXPANSION ist

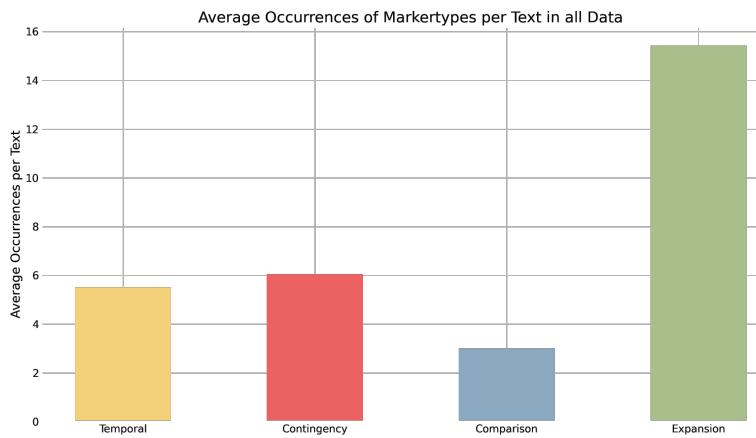


Abbildung 7.17: Durchschnittliches Vorkommen aller DM Klassen in allen Diskursarten

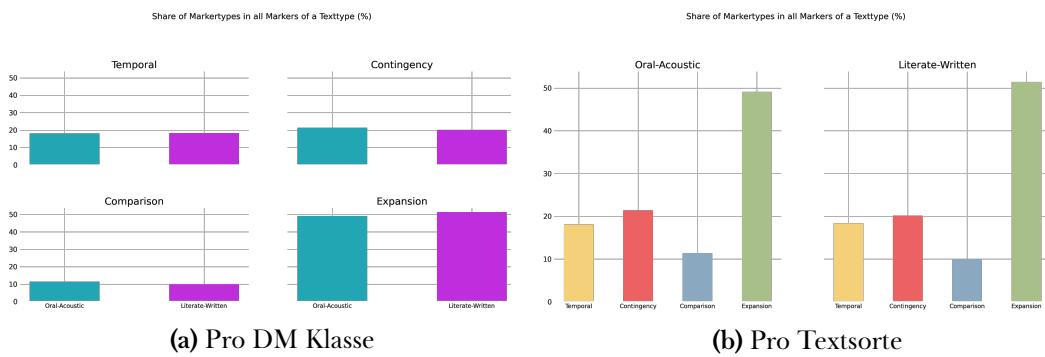


Abbildung 7.18: Prozentualer Anteil der DM Klassen an allen DM

kein statistisch signifikanter Unterschied erkennbar, doch nutzen OA Texte signifikant mehr CONTINGENCY DM als LS Texte ($p < 0.001$, EG: 0.32). Auch COMPARISON DM werden von OA Texten signifikant häufiger genutzt als von LS Texten ($p < 0.001$, EG: 0.87).

7.4.2 Vergleich zwischen Genres

Abbildung 7.19 zeigt, dass die EXPANSION DM auch bei den Genres mit durchschnittlich 152 DM pro Text den weitaus größten Teil der verwendeten DM ausmachen. TEMPORAL und CONTINGENCY liegen mit durchschnittlich 66 bzw. 71 DM etwa gleichauf, während COMPARISON DM mit durchschnittlich 46 DM pro Text den kleinsten Anteil haben.

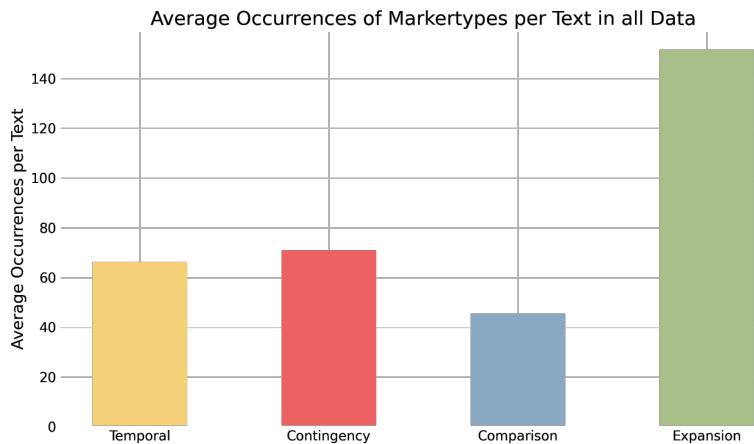


Abbildung 7.19: Durchschnittliches Vorkommen aller DM Klassen in allen Genres

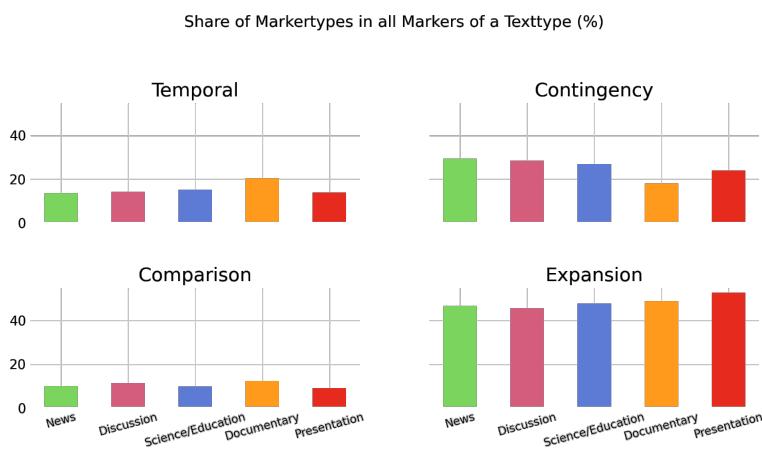


Abbildung 7.20: Prozentualer Anteil der DM Klassen an allen DM - Pro DM Klasse

TEMPORAL DM werden in allen Genres bis auf DOCUMENTARY am zweit seltensten genutzt. Ein statistisch signifikanter Unterschied lässt sich bei diesen DM nur zwischen DISCUSSION und DOCUMENTARY ($p < 0.001$, EG: 0.18) sowie zwischen DISCUSSION und PRESENTATION ($p < 0.001$, EG: 0.18) feststellen. In beiden Fällen nutzen die DISCUSSION Texte signifikant weniger TEMPORAL DM als die jeweils anderen Genres.

Für CONTINGENCY DM lässt sich ein statistisch signifikanter Unterschied zwischen PRESENTATION und jedem der anderen Genres feststellen. In jedem der Fälle nutzen PRESENTATION Texte signifikant mehr CONTINGENCY DM als die anderen Genres. Dies mag in der Natur der Präsentation liegen: CONTINGENCY DM

zeigen vor allem Schlussfolgerungen und Begründungen an.

Bei den COMPARISON DM ist zwischen fast allen Genres ein signifikanter Unterschied in der Häufigkeit der Nutzung feststellbar. Ausgenommen sind die Paare NEWS und PRESENTATION sowie SCIENCE/EDUCATION und DOCUMENTARY. In beiden Fällen scheint jeweils eine ähnliche Menge an Vergleichen mit DM signalisiert zu werden.

Für EXPANSION DM sind in vier Fällen keine statistisch signifikanten Unterschiede feststellbar. So nutzen NEWS, SCIENCE/EDUCATION und DOCUMENTARY jeweils ähnlich viele DM dieser Klasse, genau wie SCIENCE/EDUCATION und PRESENTATION.

7.4.3 Vergleich zwischen Konversationsarten

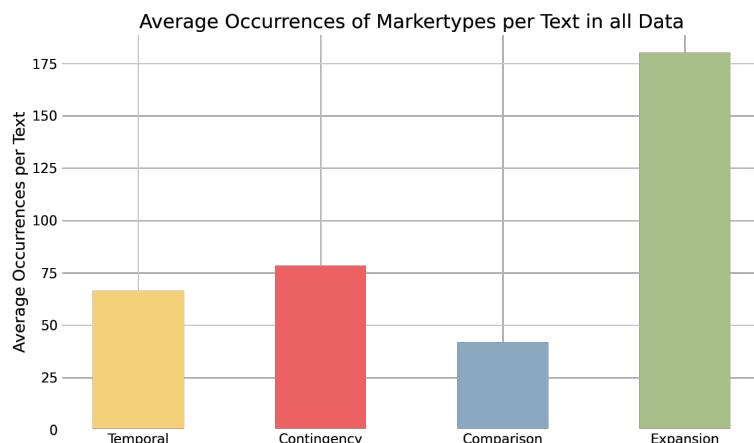


Abbildung 7.21: Durchschnittliches Vorkommen aller DM Klassen in allen Konversationsarten

Die Abbildungen 7.21 und 7.22 zeigen, dass auch bei den Konversationsarten die EXPANSION DM den weitaus größten Teil aller DM ausmachen; ca. 50% aller DM jeder Konversationsart sind dieser Klasse zuzuordnen. Ein signifikanter Unterschied in der Menge der Verwendung ist bei diesen DM zwischen fast allen Konversationsarten feststellbar. Ausgenommen sind die Paare DIALOG und REDE, sowie MONOLOG und KOOPERATIVER MONOLOG.

Den kleinsten Anteil an allen DM haben mit ca. 10% die COMPARISON DM.

Bei ihnen lässt sich zwischen allen Konversationsarten ein signifikanter Unterschied in der Häufigkeit der Nutzung feststellen.

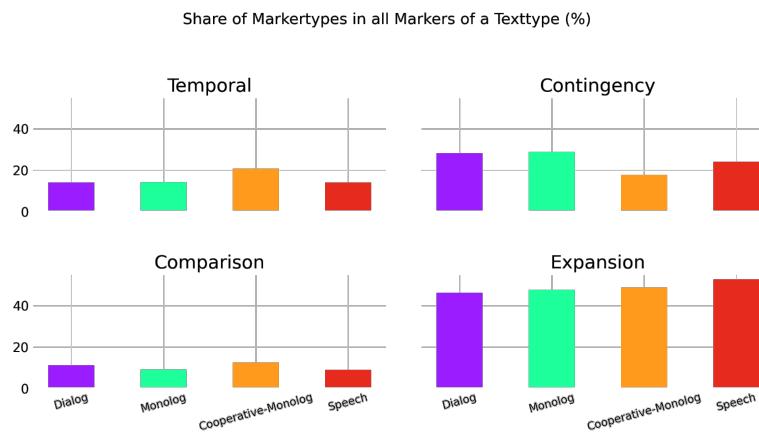


Abbildung 7.22: Prozentualer Anteil der DM Klassen an allen DM - Pro DM Klasse

So werden in DIALOGEN signifikant mehr DM dieser Klasse genutzt als in KOOPERATIVEN MONOLOGEN ($p < 0.001$, EG: 0.31), MONOLOGEN ($p < 0.001$, EG: 0.5) und REDEN ($p < 0.001$, EG: 0.62). Texte des KOOPERATIVEN MONOLOGS wiederum nutzen signifikant mehr DM als MONOLOGE ($p < 0.001$, EG: 0.27) und REDEN ($p < 0.001$, EG: 0.47). Auch MONOLOGE nutzen signifikant mehr COMPARISON DM als REDEN. REDEN nutzen insgesamt wesentlich weniger COMPARISON DM als alle anderen Konversationsarten.

Für TEMPORAL DM lassen sich zwischen den Konversationsarten keine statistisch signifikanten Unterschiede feststellen. Bei den CONTINGENCY DM lassen sich statistisch signifikante Unterschiede nur im Bezug auf REDEN feststellen. Diese nutzen mehr CONTINGENCY DM als alle anderen Konversationsarten.

7.5 Welche Diskursmarker werden innerhalb der jeweiligen Klassen besonders genutzt?

Diese Fragestellung geht auf die einzelnen Diskursmarker innerhalb der DM-Klassen ein. Für diesen Teil der Betrachtung wurden aufgrund der dafür notwendigen Menge keine statistischen Tests durchgeführt. An dieser Stelle müsste für jeden einzelnen DM ein eigener Test durchgeführt werden und die Auswer-

tung all dieser Tests würde den Umfang dieser Arbeit bei weitem übersteigen. Bei der folgenden Auswertung wird jeweils der prozentuale Anteil, den ein DM in der jeweiligen Textsorte an allen vorkommenden DM seiner Klasse hat, betrachtet. Zur Veranschaulichung wird jeweils das Diagramm mit den häufigsten TEMPORAL DM gezeigt. Die Diagramme der übrigen DM-Klassen sowie die Tabellen mit der prozentualen Zu- und Abnahme der Nutzung jedes DM pro Textsorte können in Anhang B eingesehen werden.

An dieser Stelle sei nochmals ausdrücklich darauf hingewiesen, dass das Problem der mitgezählten Homographen in dieser Arbeit ungelöst bleibt und davon ausgegangen werden muss, dass ein nicht zu unterschätzender Anteil jedes gezählten DMs tatsächlich im jeweiligen Kontext nicht die Funktion eines DMs haben könnte.

7.5.1 Vergleich zwischen Diskursarten

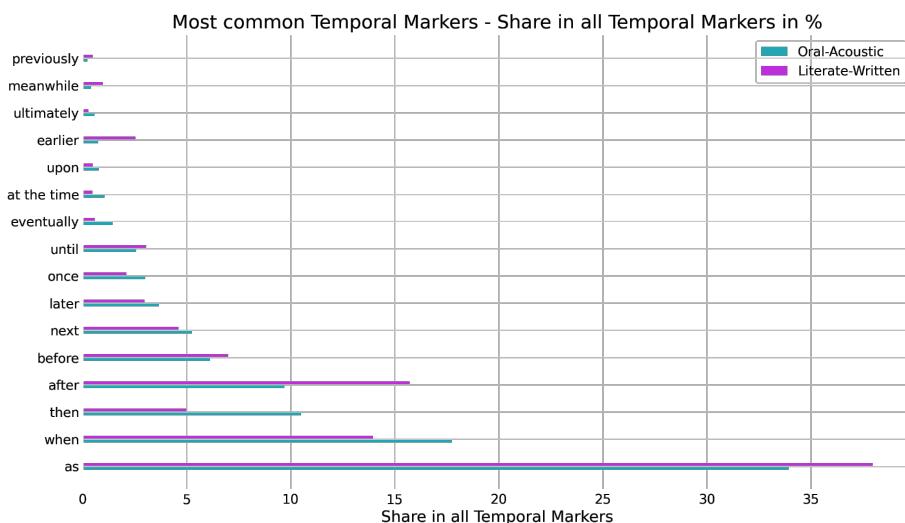


Abbildung 7.23: Diskursarten - häufigste TEMPORAL DM

Abbildung 7.23 zeigt die häufigsten TEMPORAL DM für die verschiedenen Diskursarten (vgl. auch Tabelle B.6). In OA Texten sind dies *as* (33% aller TEMPORAL DM), *when* (18%) und *then* (11%). Für LS Texte sind es *as* (38%), *after* (16%) und *when* (14%). In beiden Fällen machen allein diese DM einen Anteil von mehr als 60% aller genutzten DM dieser Klasse aus. *As* ist außerdem in beiden Fällen

der am häufigsten genutzte TEMPORAL DM. Wesentlich häufiger als in OA Texten kommen in LS Texten in dieser Klasse die DM *earlier* (+240%), *meanwhile* (+131%) und *previously* (+104%) vor. Wesentlich seltener hingegen *then* (-52%), *at that time* (-54%) und *eventually* (-59%).

Die häufigsten CONTINGENCY DM (vgl. Abbildung B.2, Tabelle B.7) in den OA Texten sind *for* (43%), *so* (26%) und *if* (14%). Für die LS Texte sind es ebenfalls *for* (66%), *if* (10%) und *so* (7%). Das sind für beide Texte über 80% aller CONTINGENCY DM. Ein auffälliger Unterschied ist hier die Nutzung von *since* und *because*. *Since* macht in den LS Texten einen wesentlich größeren Teil der CONTINGENCY DM aus als in den OA Texten (6% vs. 2%). *Because* hingegen hat in den OA Texten einen größeren Anteil als in den LS Texten (9% vs. 5%). Beide Wörter werden als CONTINGENCY DM eingesetzt, um eine Begründung zu signalisieren. CONTINGENCY DM, die in LS Texten wesentlich häufiger vorkommen als in OA Texten sind *in response to* (+348%), *since* (+237%) und *because of* (+195%). Wesentlich seltener als in OA werden in LS *because* (-42%), *now that* (-49%) und *whatever* (-63%) verwendet.

Die häufigsten COMPARISON DM (vgl. Abbildung B.3, Tabelle B.8) in OA Texten sind *but* (65%), *still* (9%) und *while* (7%). In LS Texten sind es die gleichen DM, jedoch liegt *while* (12%) vor *still* (11%). *But* steht auch hier mit 56% an Platz 1. Wesentlich häufiger als in OA Texten kommen in LS Texten die COMPARISON DM *in contrast* (+288%), *nonetheless* (+242%) und *although* (+118%) vor. Wesentlich seltener hingegen werden die DM *as though* (-47%), *on the other hand* (-55%) und *by the way* (-82%) verwendet.

Der mit Abstand häufigste EXPANSION DM (vgl. Abbildung B.4, Tabelle B.9) ist *and* (OA: 68%, LS: 64%), gefolgt von *with* (OA: 15%, LS: 18%) und *or* (OA: 7%, LS: 6%). Das sind für beide Textsorten rund 90% aller EXPANSION DM. Wesentlich häufiger als in OA Texten werden in LS Texten die EXPANSION DM *overall* (+274%), *unless* (+99%) und *as well* (+97%) verwendet. Wesentlich seltener hingegen *anyway* (-64%), *in fact* (-68%) und *finally* (-70%).

7.5.2 Vergleich zwischen Genres

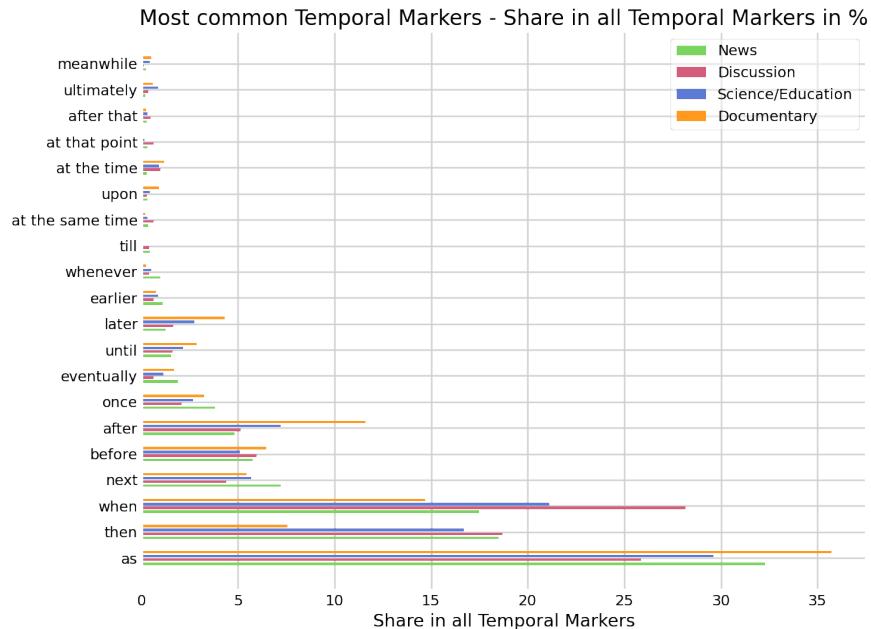


Abbildung 7.24: Genres - häufigste TEMPORAL DM

Abbildung 7.24 zeigt die häufigsten TEMPORAL DM für die verschiedenen Genres (vgl. auch Tabelle B.10). Die drei häufigsten sind *as*, *then* und *when*. Einzig DOCUMENTARY verwendet *after* wesentlich häufiger als *then*. Auffällig ist außerdem der DM *ultimately*, der in NEWS wesentlich häufiger verwendet wird als in den anderen Genres.

Die häufigsten CONTINGENCY DM (vgl. Abbildung B.6, Tabelle B.11) sind *so* (rund 38%), *for* (rund 30%) und *if* (rund 14%). Wieder unterscheidet sich DOCUMENTARY durch eine andere Verteilung der Häufigkeiten: hier ist *for* mit Abstand der am häufigsten verwendete CONTINGENCY DM (56%), gefolgt von *so* (16%) und *if* (14%). Auffällige DM sind an dieser Stelle *after all* und *as a result*. Beide werden in NEWS wesentlich seltener verwendet als in allen anderen Genres, am häufigsten jedoch in DOCUMENTARY und PRESENTATION.

Der häufigste COMPARISON DM (vgl. Abbildung B.7, Tabelle B.12) ist mit Abstand *but*. DOCUMENTARY verwendet diesen DM von allen Genres am seltens-

ten (60%), während DISCUSSION mit 74% bei der Nutzung von *but* am weitesten vorne liegt. An zweiter Stelle liegt *still* mit rund 9%. Der dritthäufigste COMPARISON DM ist nicht ganz eindeutig. Für NEWS und DISCUSSION ist es *though*, für SCIENCE/EDUCATION, DOCUMENTARY und PRESENTATION hingegen *while*. Die drei häufigsten COMPARISON DM machen in allen Genres mehr als 75% aller COMPARISON DM aus. Ein auffälliger DM dieser Klasse ist *despite*, der in den eher sachlichen Genres DOCUMENTARY, SCIENCE/EDUCATION und PRESENTATION wesentlich häufiger verwendet wird als in den übrigen Genres.

Der mit Abstand häufigste EXPANSION DM (vgl. Abbildung B.8, Tabelle B.13) ist auch bei den Genres *and*. Am seltesten wird *and* mit 65% in NEWS verwendet, am häufigsten mit 72% in DISCUSSION. Die zweit- und dritthäufigsten EXPANSION DM sind für alle Genres *with* (rund 14%) und *or* (rund 8%). Auch hier machen die drei häufigsten EXPANSION DM für jedes Genre mehr als 75% aller DM dieser Klasse aus.

7.5.3 Vergleich zwischen Konversationsarten

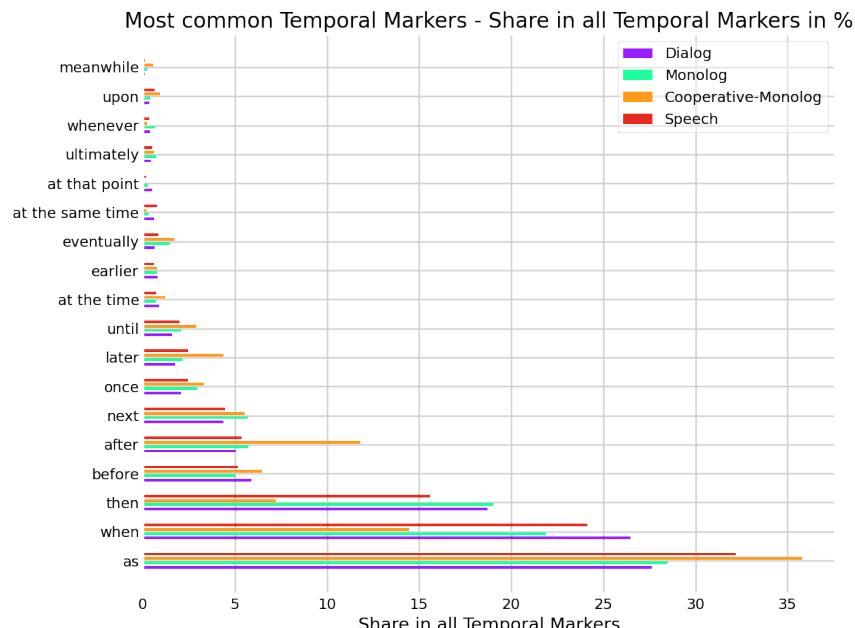


Abbildung 7.25: Konversationsarten - häufigste TEMPORAL DM

Abbildung 7.25 zeigt die häufigsten TEMPORAL DM für alle Konversationsarten (vgl. auch Tabelle B.14). Die beiden häufigsten sind für alle Konversationsarten *as* (rund 31%) und *when* (rund 22%). Am häufigsten wird *as* im KOOPERATIVEN MONOLOG genutzt (36%) und am seltesten im DIALOG (28%). Am dritthäufigsten nutzen alle Konversationsarten außer KOOPERATIVER MONOLOG den DM *then*, während im KOOPERATIVEN MONOLOG *after* Platz drei einnimmt. Auffällige TEMPORAL DM sind außerdem *eventually* und *meanwhile*, die im Dialog wesentlich seltener genutzt werden als in allen anderen Konversationsarten.

Die häufigsten CONTINGENCY DM (vgl. Abbildung B.10, Tabelle B.15) sind *so*, *for* und *because*. Im KOOPERATIVEN MONOLOG ist *for* mit 59% der häufigste CONTINGENCY DM, für alle anderen Konversationsarten ist es *so* mit rund 39%. Auffällig sind auch die CONTINGENCY DM *after all* und *in this way*. Beide werden im DIALOG wesentlich seltener benutzt als in allen anderen Konversationsarten.

Der häufigsten COMPARISON DM (vgl. Abbildung B.11, Tabelle B.16) ist in allen Konversationsarten mit Abstand *but*. Auffällig ist hierbei, dass *but* im KOOPERATIVEN MONOLOG mit 60% wesentlich seltener verwendet wird als in den anderen Konversationsarten, im DIALOG hingegen mit 75% am häufigsten. Häufig sind außerdem in allen Konversationsarten *still*, *though* und *while*. Auffällig ist wieder der Kontrast, den der KOOPERATIVE MONOLOG darstellt. So treten die größten Differenzen in der Verwendungshäufigkeit fast immer im Zusammenhang mit dieser Konversationsart auf. DM wie *although* und *however* sind mit Abstand am häufigsten in den geskripteten Texten des KOOPERATIVEN MONOLOGS zu finden. Auch der DIALOG hebt sich klar von den anderen Konversationsarten ab und nutzt DM wie *despite* wesentlich seltener.

Der häufigste EXPANSION DM (vgl. Abbildung B.12, Tabelle B.17) bleibt mit Abstand *and*. Im Vergleich am seltensten verwendet wird er im KOOPERATIVEN MONOLOG (66%), am häufigsten in der REDE (72%). An zweiter und dritter Stelle stehen in allen Konversationsarten die DM *with* und *or*. Gemeinsam machen diese drei DM in allen Konversationsarten mehr als 80% aller verwendeten EXPANSION DM aus. KOOPERATIVER MONOLOG und REDE stellen im Bezug auf die Nutzung einzelner EXPANSION DM einen auffälligen Kontrast zum DIALOG dar. DM wie *further*, *in fact* und *indeed* werden im DIALOG wesentlich seltener verwendet als in diesen beiden Konversationsarten.

8

Schlussfolgerung

Die in Kapitel 7 aufgeführten Ergebnisse lassen einige Schlüsse über die Nutzung von Diskursmarkern in verschiedenen Textsorten zu.

Generelle Verteilung

Texte mit aufgrund ihres Konzepts höherer Listenability, also ORAL-AKUSTISCHE Texte, nutzen generell mehr Diskursmarker als LITERAT-SCHRIFTLICHE Texte.

Bei den Genres der ORAL-AKUSTISCHEN Texte enthalten die geskripteten DOCUMENTARIES weniger DM als alle anderen Genres. Obwohl auch die PRESENTATION Texte geskriptet sind, spielt in ihnen die Lenkung der Aufmerksamkeit der Zuhörer eine größere Rolle als bei längeren, fokussierteren DOCUMENTARIES. Auch DOCUMENTARIES wurden zwar für Listenability konzipiert, doch sind dies längere Sendungen, die bewusst angehört werden, um sich über ein bestimmtes Thema zu informieren. Das Ziel eines TED-Talks ist hingegen, in kurzer Zeit die Aufmerksamkeit der Zuhörer zu erlangen und ihnen das Zuhören so einfach wie möglich zu machen. Diskursmarker spielen in diesem Zusammenhang offensichtlich eine Rolle.

Die Konversationsarten DIALOG und REDE ähneln sich in der Menge der verwendetet DM, möglicherweise aufgrund der Interaktion bzw. Miteinbeziehung anderer Personen. Die Listenability beider Konversationsarten ist hoch, ihre Natur ist es, Zuhörer in den Bann zu ziehen. Der KOOPERATIVE MONOLOG nutzt wesentlich weniger DM als alle anderen Konversationsarten, der MONOLOG hingegen wesentlich mehr. Da ein Monolog durch nur eine sprechende Person vorgetragen wird, ist es hier von äußerster Wichtigkeit, den Zuhörern Wegweiser an

die Hand zu geben. Sprecherwechsel sind nicht vorhanden und ohne eingreifende Diskurspartner können Gedankengänge sehr lang und kompliziert werden. Diskursmarker sind hier ein wichtiges Mittel, um Listenability zu gewährleisten.

Position im Dokument

ORAL-AKUSTISCHE Texte nutzen am Textanfang und -ende mehr Diskursmarker als LITERAT-SCHRIFTLICHE, diese nutzen dafür in der Mitte mehr DM als ORAL-AKUSTISCHE. Interaktionslastigere Genres nutzen am Anfang und in der Mitte des Textes mehr DM als andere, die sachlicheren Genres dafür eher am Ende. Auch bei den Konversationsarten sind es die interaktiveren, die am Anfang und in der Mitte des Textes mehr DM nutzen als die anderen. Es scheint also vom Ziel des Diskurses abzuhängen, an welcher Stelle DM eingesetzt werden. Soll Interaktion erfolgen, werden sie eher am Anfang eingesetzt, ansonsten eher am Ende, wenn die Konzentration der Zuhörer vielleicht schon etwas nachlässt.

Position im Satz

Am Satzanfang setzen ORAL-AKUSTISCHE Texte mehr DM ein als LITERAT-SCHRIFTLICHE. Diese wiederum nutzen in der Mitte mehr DM als ORAL-AKUSTISCHE Texte, während sich am Satzende kein Unterschied mehr feststellen lässt. Auch der DIALOG setzt am Satzanfang mehr auf DM als alle anderen Konversationsarten.

Verwendete Diskursmarker Klassen

EXPANSION DM machen für alle Textsorten den weitaus größten Teil aller DM aus. Diese Klasse enthält zum einen die meisten DM, zum anderen gehört ihr der extrem häufig verwendete DM *and* an. Für TEMPORAL und EXPANSION lassen sich keine signifikanten Unterschiede feststellen, doch nutzen ORAL-AKUSTISCHE Texte mehr COMPARISON und CONTINGENCY DM als LITERAT-SCHRIFTLICHE Texte.

Verwendete Diskursmarker innerhalb der Klassen

Die meistgenutzten DM sind *as* (TEMPORAL), *and* (EXPANSION), *but* (COMPARISON), *so* und *for* (beide CONTINGENCY). *And*, *but* und *so* sind ebenfalls die von Fraser [2009] als *primary markers* bezeichneten DM, die repräsentativ für ihre jeweilige

Klasse stehen. Es lässt sich feststellen, dass umgangssprachlichere DM wie *whatever* und *by the way* eher in ORAL-AKUSTISCHEN Texten und dann auch vermehrt in improvisierten, freien Formen wie DIALOG und DISCUSSION benutzt werden. Etwas gehobenere Ausdrücke wie *despite* und *since* (mit begründender Bedeutung) hingegen werden eher in LITERAT-SCHRIFTLICHEN Texten verwendet. Auch die geskripteten Formen der ORAL-AKUSTISCHEN Texte nutzen eher gehobenere DM als die improvisierten.

Zusammenfassung

Für die Verbesserung der Listenability bedeutet dies, dass korrekt eingesetzte Diskursmarker hilfreich sein können. Über das akustische Medium wahrgenommene Texte profitieren von Wegweisern, die das Verständnis der Diskursbeziehungen verbessern. Jedoch sollte darauf geachtet werden, welchem Genre oder welcher Konversationsart der Text angehört. Ist er interaktiv und bezieht das Publikum mit ein, werden mehr Diskursmarker eingesetzt, außerdem kommen sie früher im Text vor. Geht es um sachlichere Themen und kann man sicher sein, dass die Zuhörer sich bewusst über ein Thema informieren wollen, reichen weniger Diskursmarker aus und sie werden eher am Ende des Textes gebraucht, wenn die Konzentration aufrecht erhalten werden soll.

Demzufolge scheint es beispielsweise durchaus sinnvoll, Empfehlungen zur Nutzung von Diskursmarkern in den Funktionsumfang von Schreibassistenz Programmen mit einzubeziehen. Diese Programme könnten dann je nach Zielmedium, Genre oder Konversationsart basierend auf den in dieser Arbeit präsentierten Ergebnissen zu der Nutzung von mehr oder weniger Diskursmarkern raten.

9

Ausblick

Aus der Betrachtung von Diskursmarkern in ORAL-AKUSTISCHEM und LITERATURSCHRIFTLICHEM Diskurs ergeben sich einige neue Fragen, die in dieser Bachelorarbeit nicht mehr abgedeckt werden konnten und Potential für weitere Forschungsansätze bieten.

Ein interessanter Aspekt sind die Positionen der einzelnen Diskursmarker-Klassen im Satz oder Text. Werden beispielsweise TEMPORAL DM immer an der gleichen Stelle eingesetzt? Lässt sich vielleicht sogar von der Satzposition auf die Funktion eines DM schließen und kann dies helfen, DM mit mehreren möglichen Bedeutungen je nach Satzposition in einer der Klassen einzuordnen?

Ein genaueres Bild als das hier präsentierte ergäbe sich zudem, wenn Homographen effektiv aussortiert werden könnten, beispielsweise nach der bisherigen Eliminierung der ungültigen Matches. Dieser Schritt hängt jedoch auch stark von der weiteren sprachwissenschaftlichen Forschung zu Diskursmarkern ab. Solange nicht eindeutig definiert werden kann, was ein Diskursmarker ist, bleibt die Betrachtung schwierig. In diesem Zusammenhang wäre auch die Analyse auf Grundlage einer alternativen Liste von Diskursmarkern, beispielsweise den von Fraser [2009] vorgeschlagenen DM und Bedeutungsgruppen, interessant.

Letztendlich wäre es von großer Hilfe, die Listenability aller betrachteten Texte mit einer geeigneten Methode zu messen und anschließend zu prüfen, ob tatsächlich ein Zusammenhang zwischen Listenability und der Nutzung von Diskursmarkern besteht.

A

Tabellen

Anmerkung: Alle hier aufgeführten Tabellen liegen auch im Repository dieser Arbeit:

<https://git.webis.de/code-teaching/theses/thesis-sacher/-/tree/master>

| COMPARISON | | CONTINGENCY | | EXPANSION | | TEMPORAL | |
|-------------------|-------|----------------|--------|-------------------|--------|------------------|--------|
| Word | ER | Word | ER | Word | ER | Word | ER |
| although | 0.3% | accordingly | 0.0% | additionally | 0.0% | after | 0.35% |
| as though | 40.0% | after all | 0.0% | also | 0.06% | after that | 0.0% |
| but | 2.81% | as a result | 0.0% | alternatively | 0.0% | afterward | 0.0% |
| by comparison | 0.0% | as a result of | 0.0% | and | 3.17% | as | 29.74% |
| by contrast | 0.0% | as long as | 33.33% | anyway | 0.0% | as soon as | 0.0% |
| by the way | 0.0% | because | 0.0% | as an alternative | 0.0% | at that point | 0.0% |
| conversely | 0.0% | because of | 0.0% | as if | 31.25% | at the same time | 0.0% |
| despite | 0.0% | consequently | 0.0% | as well | 0.0% | at the time | 0.0% |
| even if | 0.0% | for | 0.0% | aside from | 0.0% | before | 0.0% |
| even so | 0.0% | given | 0.0% | besides | 5.26% | before and after | 0.0% |
| even though | 0.0% | given that | 0.0% | either or | 0.0% | by then | 0.0% |
| however | 0.41% | hence | 0.0% | else | 0.0% | earlier | 6.67% |
| in contrast | 0.0% | if | 4.01% | essentially | 0.0% | eventually | 0.0% |
| in spite of | 0.0% | if and when | 0.0% | except | 0.0% | everytime | 0.0% |
| instead of | 0.0% | if then | 0.0% | except that | 0.0% | in turn | 36.67% |
| nevertheless | 0.0% | in case | 0.0% | finally | 37.5% | later | 1.1% |
| nonetheless | 3.7% | in response to | 0.0% | for example | 0.0% | meantime | 13.33% |
| on the contrary | 0.0% | in this way | 0.0% | for instance | 0.0% | meanwhile | 51.3% |
| on the one hand | 0.0% | inssofar as | 0.0% | for one | 0.0% | next | 14.29% |
| on the other hand | | | | | | | |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Comparison | | Contingency | | Expansion | | Temporal | |
|-------------------|--------|-----------------|--------|-------------------------|--------|----------------|--------|
| Word | ER | Word | ER | Word | ER | Word | ER |
| on the other hand | 2.7% | irrespective of | 0.0% | for one thing | 0.0% | once | 4.76% |
| regardless | 0.0% | lest | 50.0% | further | 0.0% | previously | 4.08% |
| still | 1.58% | now that | 31.82% | furthermore | 0.0% | simultaneously | 0.0% |
| though | 0.0% | since | 47.83% | in addition | 0.0% | then | 6.76% |
| whereas | 0.0% | so | 0.0% | in addition to | 0.0% | thereafter | 0.0% |
| while | 33.93% | so that | 0.0% | in any case | 0.0% | till | 0.0% |
| yet | 2.97% | thereby | 0.0% | in any event | 0.0% | ultimately | 22.22% |
| | | therefore | 0.0% | in essence | 0.0% | until | 12.96% |
| | | thus | 0.0% | in fact | 7.32% | upon | 0.0% |
| | | whatever | 0.0% | in other words | 0.0% | when | 19.82% |
| | | when/then | 0.0% | in particular | 0.0% | when and if | 0.0% |
| | | | | in short | 0.0% | whenever | 0.0% |
| | | | | in sum | 0.0% | | |
| | | | | in the end | 44.44% | | |
| | | | | indeed | 4.81% | | |
| | | | | instead | 2.68% | | |
| | | | | likewise | 0.0% | | |
| | | | | moreover | 0.0% | | |
| | | | | much as | 50.0% | | |
| | | | | neither nor | 33.33% | | |
| | | | | nor | 3.23% | | |
| | | | | not only/but | 0.0% | | |
| | | | | or | 3.06% | | |
| | | | | otherwise | 0.0% | | |
| | | | | overall | 0.0% | | |
| | | | | particularly | 0.0% | | |
| | | | | plus | 0.0% | | |
| | | | | quite the contra- ry | 0.0% | | |
| | | | | rather | 23.53% | | |
| | | | | rather than | 0.0% | | |
| | | | | separately | 0.0% | | |
| | | | | similarly | 0.0% | | |
| | | | | specifically | 0.0% | | |
| | | | | unless | 1.05% | | |

Fortsetzung auf nächster Seite ...

| Comparison | | Contingency | | Expansion | | Temporal | |
|------------|----|-------------|----|-----------|------|----------|----|
| Word | ER | Word | ER | Word | ER | Word | ER |
| | | | | with | 0.0% | | |
| | | | | without | 0.0% | | |

Tabelle A.1: Tabelle aller betrachteten Diskursmarker inkl. Error Rate (ER) für die entsprechende Kategorie

| Titel | Sprache | Kostenlos | Art |
|--|----------|-----------|-------------|
| 2002 Rich Transcription Broadcast News and Conversational Telephone Speech | Englisch | Nein | Spontan |
| AMI Corpus | Englisch | Ja | Spontan |
| Boston University Radio Speech Corpus | Englisch | Nein | Nachrichten |
| Buckeye Corpus of conversational speech | Englisch | Nein | Interviews |
| Datenbank für Gesprochenes Deutsch | Deutsch | Ja | Sammlung |
| RadioTalk: A large-scale corpus of talk radio transcripts | Englisch | Ja | Talk Radio |
| Spotify Podcast Dataset | Englisch | Ja | Podcasts |
| TED-LIUM 3 | Englisch | Ja | TED-Talks |

Tabelle A.2: Liste der in Betracht gezogenen akustischen Corpora

| Show | Tags | Snippets |
|---|-----------------------------------|-----------|
| ABC News Perspective | news(ish) | 4 570 |
| Alabama's Morning News with JT | interviews, news(ish) | 72 336 |
| America's Morning News (Washington Times) | news(ish), political | 133 411 |
| Arizona's Morning News | news(ish) | 133 860 |
| Atlanta's Morning News | news(ish) | 133 860 |
| BBC World Service | interviews, news(ish), political | 5 323 632 |
| Ben Shapiro | entertainment, opinion, political | 46 993 |
| Big Picture Science | entertainment, science, stories | 9 589 |
| Billings Morning News | news(ish), political | 133 870 |
| CBS News Weekend Roundup | news(ish), opinion, political | 4 420 |
| Columbia s Morning News | news(ish) | 57 398 |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Tags | Snippets |
|--|--|-----------|
| Crosscurrents from KALW News | entertainment, news(ish), political | 13 721 |
| Dr Drew Midday Live with Mike Catherwood | calls, entertainment, news(ish), political | 101 856 |
| First Morning News | news(ish) | 18 122 |
| Gary Sadlemeyer and KFAB's Morning News | interviews, news(ish), political | 85 387 |
| Here and Now (PRI) | interviews, news(ish), political | 1 112 012 |
| Houston's Morning News | news(ish) | 57 893 |
| In Deep with Angie Coiro | entertainment, interviews, political | 4 470 |
| Jeff Caplan's Afternoon News | entertainment, news(ish), political | 185 870 |
| KFAB Morning News, Weather & Markets | news(ish) | 3 556 |
| KOGO Weekend News Edition | news(ish) | 6 984 |
| KOGO's Evening News | news(ish) | 18 143 |
| Leslie Marshall | entertainment, opinion, political | 178 622 |
| McIntyre In the Morning | entertainment, news(ish), political | 156 532 |
| Mid-Morning News Hour (WBSM) | news(ish) | 33 047 |
| Morning Edition (NPR) | entertainment, news(ish), political | 4 028 929 |
| MORNING NEWS | news(ish) | 74 797 |
| Nashville's Morning News | news(ish) | 124 348 |
| National Native News | news(ish), opinion, political | 4 226 |
| Newshour Extra | news(ish) | 6 833 |
| PBS NewsHour | live, news(ish), opinion, political | 54 774 |
| PRI's The World | news(ish), opinion, political | 564 317 |
| Sam Nation | entertainment, opinion, political | 32 399 |
| Saturday Morning News wDave Russell | news(ish) | 6 898 |
| The Daily | entertainment, news(ish), political | 74 741 |
| The Thom Hartmann Program | entertainment, news(ish), opinion, political | 275 018 |
| Utah's Morning News | news(ish), political | 135 039 |
| Utah's Noon News | news(ish), political | 22 539 |
| WBRC Fox 6 News at 10 | news(ish) | 11 465 |
| WBZ Afternoon News | news(ish) | 140 823 |
| WBZ Midday News | news(ish) | 125 111 |
| WBZ Morning News | news(ish) | 149 703 |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Tags | Snippets |
|----------------------------|--------------------|----------|
| Weekly NH News Roundup | news(ish) | 8 189 |
| WFAW News | news(ish), opinion | 104 959 |
| Wisconsin's Afternoon News | news(ish) | 140 295 |
| Wisconsin's Morning News | news(ish) | 142 257 |
| WMT News Hour | news(ish) | 3 183 |
| WORT Local News | news(ish) | 14 234 |
| WPRO Morning News | news(ish) | 150 074 |

Tabelle A.3: Liste der potentiell relevanten Shows des RadioTalk Corpus

ANHANG A. TABELLEN

Von: Doug Beeferman dougb5@mit.edu
Betreff: Re: Punctuation in the RadioTalk Corpus
Datum: 15. Juni 2020 um 20:24
An: Johanna Sacher johanna.sacher@uni-weimar.de, William Walker Brannon wbrannon@mit.edu, dkroy@media.mit.edu

Hi Johanna,

Thanks for your interest in using the corpus. I don't have a great answer for you, since as you point out the punctuation in the source audio is not verbalized. Restoring the punctuation to such a speech recognition transcript is a separate post-processing step that we haven't done within the scope of the RadioTalk work. But many people have looked into this over the years – this ICASSP paper from last year and its citations might make for a good starting point:
https://ieeexplore.ieee.org/abstract/document/8682418?casa_token=DCFAzH9IU-QAAAAA:HHiovqSe1gWKdoS6gskR0rzCYH_ueQadrWls1KGHw2sdXLtpUuaVvdRIOnX3RejQprq9Gwl

In the RadioTalk corpus, the snippet boundaries are decided purely based on audio features, using the output of the same LIUM diarization package that identifies speaker turns. This algorithm puts boundaries wherever the speaker changes, as well as in the middle of long speaker turns. When it chops up a long speaker turn, it prefers chopping within silence, which are often but not always sentence boundaries. See the section "Segmentation based on Viterbi decoding" on <https://projets-lium.univ-lemans.fr/spkdiarization/quick-start/> for a more precise description.

For our own linguistic analyses of the corpus, we have usually treated snippets as "noisy" sentences, good enough for topic analysis but admittedly error-prone if you rely on the parse of the sentence in some way in your downstream application.

Thanks,

Doug

From: Johanna Sacher <johanna.sacher@uni-weimar.de>
 Sent: Monday, June 15, 2020 9:30 AM
 To: Doug Beeferman; William Walker Brannon; dkroy@media.mit.edu
 Subject: Punctuation in the RadioTalk Corpus

Dear Mr. Beeferman, Mr. Brannon and Mr. Roy,

I am writing to you because I have a question regarding the RadioTalk Corpus. I am currently working on my Bachelor Thesis which is part of the research project „Conversational News“ at Bauhaus University Weimar; we would love to use part of your corpus as a base for analysing whether certain written news are suitable for being read aloud.
 While analysing which shows in your data are news related and thus usable for us I noticed that there is no punctuation in the transcripts of the audio. I guess this is probably due to the fact that they are automatic transcripts and punctuation is hard to transcribe? But as we need the data to analyse things like the length of a sentence, punctuation would be really helpful. In your paper I was unable to find anything concerning this problem and wanted to ask if there would be any way for me to add some punctuation? I tried simply inserting full-stops between all snippets when putting them back together (I tried to retrieve the coherent texts of each audio chunk by putting the respective snippets in the right order) but while this worked sometimes, other times it just cut a sentence in half. Maybe knowing how the snippets were generated could help me?
 Any suggestion would be highly appreciated.

Thank you very much and kind regards
 Johanna Sacher

Abbildung A.1: E-Mail Konversation mit Doug Beeferman über die fehlende Interpunktions des RadioTalk Corpus

| Show | Genre | Konversationsart |
|---|-------------------|------------------|
| #5 Things | Discussion | Dialog |
| 10 Awesome New Inventions in USA that You'll Never Hear About . | Science/Education | Dialog |
| 22 Yarns with Gaurav Kapur | Discussion | Dialog |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Genre | Konversationsart |
|--|-------------------|------------------|
| 3Bawls Basketball Podcast | Discussion | Dialog |
| A Tap On The Wrist | Documentary | Dialog |
| A-Pod Cast For Killer Whales | Science/Education | Monolog |
| ABA Ultimate Showdown! | Science/Education | Dialog |
| AI Ireland Podcast | Documentary | Dialog |
| AllDatNoise Podcast | Discussion | Dialog |
| American Girls | Documentary | Dialog |
| Anarchopac | Science/Education | Dialog |
| Ancient Gods and Demons | Documentary | Monolog |
| Anthropological Theory: A podcast created by anthropology students | Science/Education | Dialog |
| APUSH Deep Dives | Science/Education | Monolog |
| Assassinations | Documentary | Coop.-Monolog |
| Authentic Biochemistry | Science/Education | Monolog |
| Autonomous Cars with Marc Hoag | News | Monolog |
| Beautiful Humans: The Social ChangeCast | Discussion | Dialog |
| Behind the Bull | Discussion | Dialog |
| Bigger Than Us | Science/Education | Dialog |
| Biology for Bastards | Science/Education | Monolog |
| Biology with Hash | Science/Education | Monolog |
| BirdsoverBoys Podcast | News | Dialog |
| Black History Buff Podcast | Documentary | Monolog |
| Born Free Podcast | Documentary | Dialog |
| Brainscape | Science/Education | Dialog |
| Brass Tacks History | Science/Education | Dialog |
| Breaking Math Podcast | Science/Education | Monolog |
| Browns part ways with general manager John Dorsey | News | Monolog |
| Bruins Take by CCSN | Discussion | Monolog |
| Burks at Loyola - The Podcast | Science/Education | Monolog |
| Carlos and Marc2019s Podcast | Discussion | Dialog |
| Carpool Chemistry | Science/Education | Monolog |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Genre | Konversationsart |
|------------------------------------|-------------------|------------------|
| CBD Deep Dive | Discussion | Dialog |
| Climate Scientists | Discussion | Dialog |
| Coast 2 Coast Fantasy Football | Discussion | Dialog |
| Coffee Tea and AI | News | Dialog |
| Con Artists | Documentary | Coop.-Monolog |
| Conspiracy or Just a Coincidence? | Documentary | Monolog |
| Conspiracy Theories | Documentary | Coop.-Monolog |
| Conviction | Documentary | Coop.-Monolog |
| Could We Ever... | Science/Education | Dialog |
| crack-a-can | News | Dialog |
| Crimes of Passion | Documentary | Coop.-Monolog |
| Crimetown | Documentary | Coop.-Monolog |
| Criminal Behaviorology | Science/Education | Dialog |
| Criminal Prints | Documentary | Monolog |
| Cults | Documentary | Coop.-Monolog |
| Dark Geometry | Science/Education | Dialog |
| Dictators | Documentary | Coop.-Monolog |
| Dose Makes The Poison: The Toxcast | Science/Education | Dialog |
| Double Helix History | Science/Education | Dialog |
| Economics In Ten | Documentary | Dialog |
| EMT A to Z | Discussion | Monolog |
| Espionage | Documentary | Coop.-Monolog |
| Ethan and Mike's Sports Show | Discussion | Dialog |
| Evolving with Mr. V | News | Monolog |
| Extraterrestrial | Documentary | Coop.-Monolog |
| Famous Fates | Documentary | Coop.-Monolog |
| Fan Wonderland | Discussion | Dialog |
| Female Criminals | Documentary | Coop.-Monolog |
| Fictorical | Documentary | Dialog |
| Flip the Switch | Discussion | Dialog |
| For The Culture | News | Dialog |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Genre | Konversationsart |
|---|-------------------|------------------|
| For Whom the Cell Tolls: Biology, Cancer, and Other Stories About Life | Science/Education | Monolog |
| Forensic Psychology | Science/Education | Dialog |
| From the Earth to the Moon: A Retrospective Podcast on The Apollo Program | Documentary | Dialog |
| Germs and Worms | Science/Education | Dialog |
| GONE | Documentary | Coop.-Monolog |
| Great Women of Business | Documentary | Coop.-Monolog |
| Haunted Places | Documentary | Coop.-Monolog |
| Historia Obscura | Documentary | Coop.-Monolog |
| Historical Figures | Documentary | Coop.-Monolog |
| History Does You | Science/Education | Dialog |
| History Under Your Feet | Documentary | Monolog |
| Hostage | Documentary | Coop.-Monolog |
| I Believe USMNT Podcast | News | Dialog |
| Immersive Chemistry Podcast | Science/Education | Dialog |
| Improbable Research | Science/Education | Dialog |
| Indirect Message | Science/Education | Monolog |
| InfoCast 5000 | Science/Education | Dialog |
| Jemele Hill is Unbothered | Discussion | Dialog |
| Just Talk'n Sports | News | Dialog |
| Kingpins | Documentary | Coop.-Monolog |
| Latest news update. | News | Monolog |
| Launch Escape: Space Podcast | Science/Education | Monolog |
| Let's Talk Sport | News | Dialog |
| Long Lost: An Investigative History Series | Documentary | Coop.-Monolog |
| Love Will Tear Us Apart | Documentary | Coop.-Monolog |
| Mardi Gras Beyond The Beads | Science/Education | Dialog |
| Medical Mysteries | Documentary | Coop.-Monolog |
| Mr. America | Discussion | Monolog |
| Natural Disasters | Documentary | Coop.-Monolog |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Genre | Konversationsart |
|-------------------------------------|-------------------|------------------|
| Nfl weekly news | News | Monolog |
| NinjAnatomy | Science/Education | Monolog |
| Not Guilty | Documentary | Coop.-Monolog |
| Oldest Stories | Documentary | Monolog |
| One Nil Football Podcast | Discussion | Dialog |
| Parcast Presents: March Mysteries | Documentary | Coop.-Monolog |
| Podkarsh with H | Discussion | Dialog |
| Political Scandals | Documentary | Coop.-Monolog |
| Pwns Media | Discussion | Dialog |
| Queen V: The Life of Queen Victoria | Documentary | Coop.-Monolog |
| Red Tail Talk Show | News | Dialog |
| Revving Up History | Science/Education | Monolog |
| Run The Fade | Discussion | Dialog |
| Science Denial | Science/Education | Monolog |
| Science Vs | Science/Education | Monolog |
| Secret Societies | Documentary | Coop.-Monolog |
| Serial Killers | Documentary | Coop.-Monolog |
| Sports Criminals | Documentary | Coop.-Monolog |
| Survival | Documentary | Coop.-Monolog |
| Technology Labs | News | Dialog |
| The Blood Evidence | Documentary | Monolog |
| The Cut on Tuesdays | Discussion | Dialog |
| The Cyclo Edition | Science/Education | Dialog |
| The Dark Side Of | Science/Education | Coop.-Monolog |
| The Fantastic History Of Food | Documentary | Monolog |
| The Genealogical Gentleman | Science/Education | Monolog |
| The Hand-off | Discussion | Dialog |
| The Hidden Tigers of Southeast Asia | Documentary | Coop.-Monolog |
| The Miami Corridor Podcast | News | Monolog |
| The Michael Decon Program | News | Dialog |
| The Mister(ess) of Pop Culture | Discussion | Monolog |

Fortsetzung auf nächster Seite ...

ANHANG A. TABELLEN

| Show | Genre | Konversationsart |
|--|-------------------|------------------|
| The Nerd ² Pod | Discussion | Dialog |
| The Whole 9 Yards | Discussion | Dialog |
| Today in True Crime | Documentary | Coop.-Monolog |
| TwinTalkYanks | News | Dialog |
| Tyrone Eagle Eye News Podcasts | News | Dialog |
| Uncivil | Documentary | Coop.-Monolog |
| Undone | Documentary | Coop.-Monolog |
| Unexplained Mysteries | Documentary | Coop.-Monolog |
| Up-to-the-minute NBA news. | News | Monolog |
| Usa Today Sports news | News | Dialog |
| Villains | Documentary | Coop.-Monolog |
| Waterloo Warriors Armchair Quarterback | News | Dialog |
| Weakly Update | News | Dialog |
| Weird Historian | Science/Education | Monolog |
| What's Trending? | Discussion | Dialog |
| Without A Country | Discussion | Dialog |

Tabelle A.4: Liste aller relevanten Spotify Podcast Shows mit Genre und Konversationsart

B

Diagramme und Statistiken

Anmerkung: Unter **Data** aufgeführt sind die jeweils miteinander verglichenen Textsorten. Die Spalte < **0.005** gibt an, ob der gemessene Unterschied statistisch signifikant ist. Ist die Effektgröße (**EG**) positiv bedeutet dies, dass die je erstgenannte Textsorte mehr DM enthält als die zweitgenannte. Wurden mehrere Statistiken in einer Tabelle zusammengefasst, ist in der Spalte **Statistic** angegeben, um welche es sich jeweils handelt.

- **SB** - Sentence Begin
- **SM** - Sentence Middle
- **SE** - Sentence End
- **DB** - Document Begin
- **DM** - Document Middle
- **DE** - Document End

| Data | < 0.05 | P-Value | EG |
|---------------------------------|--------|---------|--------|
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | 1.636 |
| News, Discussion | ✗ | 0.168 | 0.17 |
| News, Science/Education | ✗ | 0.969 | -0.004 |
| News, Documentary | ✓ | < 0.001 | 0.864 |
| News, Presentation | ✗ | 0.501 | 0.062 |
| Discussion, Science/Education | ✗ | 0.065 | -0.163 |
| Discussion, Documentary | ✓ | < 0.001 | 0.689 |
| Discussion, Presentation | ✗ | 0.178 | -0.085 |
| Science/Education, Documentary | ✓ | < 0.001 | 0.797 |
| Science/Education, Presentation | ✗ | 0.294 | 0.063 |
| Documentary, Presentation | ✓ | < 0.001 | -0.667 |
| Dialog, Monolog | ✓ | 0.006 | -0.215 |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | 0.754 |
| Dialog, Speech | ✗ | 0.442 | -0.042 |
| Monolog, Cooperative Monolog | ✓ | < 0.001 | 0.992 |
| Monolog, Speech | ✓ | 0.008 | 0.15 |
| Cooperative Monolog, Speech | ✓ | < 0.001 | -0.697 |

Tabelle B.1: Zu Abschnitt 7.1: Vergleich der prozentualen Anteile der DM an allen Wörtern der Textsorten

| Data | < 0.05 | P-Value | EG |
|--------------------------------|--------|---------|--------|
| Spotify, Literate-Written | ✓ | < 0.001 | 0.347 |
| News, Discussion | ✗ | 0.051 | -0.018 |
| News, Science/Education | ✗ | 0.117 | 0.015 |
| News, Documentary | ✓ | < 0.001 | 0.047 |
| Discussion, Science/Education | ✓ | < 0.001 | 0.033 |
| Discussion, Documentary | ✓ | < 0.001 | 0.064 |
| Science/Education, Documentary | ✓ | < 0.001 | 0.032 |
| Dialog, Monolog | ✓ | < 0.001 | 0.036 |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | 0.066 |
| Monolog, Cooperative Monolog | ✓ | < 0.001 | 0.03 |

Tabelle B.2: Zu Abschnitt 7.1: Vergleich der DM pro Satz

| Data | < 0.05 | P-Value | EG | Statistic |
|---------------------------------|--------|---------|--------|-----------|
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | 0.111 | DB |
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | -0.136 | DM |
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | 0.057 | DE |
| News, Discussion | ✗ | 0.213 | -0.152 | DB |
| News, Science/Education | ✗ | 0.213 | 0.146 | DB |
| News, Documentary | ✗ | 0.033 | 0.217 | DB |
| News, Presentation | ✗ | 0.274 | -0.092 | DB |
| Discussion, Science/Education | ✓ | < 0.001 | 0.297 | DB |
| Discussion, Documentary | ✓ | < 0.001 | 0.361 | DB |
| Discussion, Presentation | ✗ | 0.643 | 0.0279 | DB |
| Science/Education, Documentary | ✗ | 0.214 | 0.075 | DB |
| Science/Education, Presentation | ✓ | < 0.001 | -0.207 | DB |
| Documentary, Presentation | ✓ | < 0.001 | -0.262 | DB |
| News, Discussion | ✗ | 0.499 | -0.085 | DM |
| News, Science/Education | ✗ | 0.099 | 0.201 | DM |
| News, Documentary | ✗ | 0.263 | 0.131 | DM |
| News, Presentation | ✗ | 0.404 | 0.08 | DM |
| Discussion, Science/Education | ✓ | < 0.001 | 0.316 | DM |
| Discussion, Documentary | ✓ | < 0.001 | 0.247 | DM |
| Discussion, Presentation | ✓ | 0.008 | 0.163 | DM |
| Science/Education, Documentary | ✗ | 0.168 | -0.091 | DM |
| Science/Education, Presentation | ✗ | 0.078 | -0.096 | DM |
| Documentary, Presentation | ✗ | 0.371 | -0.027 | DM |
| News, Discussion | ✗ | 0.085 | 0.23 | DE |
| News, Science/Education | ✓ | 0.007 | -0.34 | DE |
| News, Documentary | ✓ | 0.008 | -0.33 | DE |
| News, Presentation | ✗ | 0.859 | -0.018 | DE |
| Discussion, Science/Education | ✓ | < 0.001 | -0.712 | DE |
| Discussion, Documentary | ✓ | < 0.001 | -0.753 | DE |
| Discussion, Presentation | ✓ | < 0.001 | -0.251 | DE |
| Science/Education, Documentary | ✗ | 0.563 | 0.039 | DE |
| Science/Education, Presentation | ✓ | < 0.001 | 0.317 | DE |
| Documentary, Presentation | ✓ | < 0.001 | 0.306 | DE |
| Dialog, Monolog | ✓ | 0.006 | 0.218 | DB |

Fortsetzung auf nächster Seite ...

| Data | < 0.05 | P-Value | EG | Statistic |
|------------------------------|--------|---------|--------|-----------|
| Dialog, Cooperative Monolog | ✓ | < 0.001 | 0.344 | DB |
| Dialog, Speech | ✗ | 0.96 | -0.003 | DB |
| Monolog, Cooperative Monolog | ✗ | 0.069 | 0.111 | DB |
| Monolog, Speech | ✓ | 0.001 | -0.174 | DB |
| Cooperative Monolog, Speech | ✓ | < 0.001 | -0.27 | DB |
| Dialog, Monolog | ✓ | < 0.001 | 0.303 | DM |
| Dialog, Cooperative Monolog | ✓ | 0.003 | 0.192 | DM |
| Dialog, Speech | ✓ | 0.016 | 0.129 | DM |
| Monolog, Cooperative Monolog | ✗ | 0.03 | -0.142 | DM |
| Monolog, Speech | ✗ | 0.017 | -0.13 | DM |
| Cooperative Monolog, Speech | ✗ | 0.501 | -0.021 | DM |
| Dialog, Monolog | ✓ | < 0.001 | -0.562 | DE |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | -0.584 | DE |
| Dialog, Speech | ✓ | 0.002 | -0.165 | DE |
| Monolog, Cooperative Monolog | ✗ | 0.332 | 0.064 | DE |
| Monolog, Speech | ✓ | < 0.001 | 0.327 | DE |
| Cooperative Monolog, Speech | ✓ | < 0.001 | 0.306 | DE |

Tabelle B.3: Zu Abschnitt 7.2: Vergleich der DM pro Position im Text

| Data | < 0.05 | P-Value | EG | Statistic |
|--------------------------------|--------|---------|--------|-----------|
| Spotify, Literate-Written | ✓ | < 0.001 | 0.381 | SB |
| Spotify, Literate-Written | ✓ | < 0.001 | -0.337 | SM |
| Spotify, Literate-Written | ✗ | 0.313 | 0.006 | SE |
| News, Discussion | ✓ | < 0.001 | 0.545 | SB |
| News, Science/Education | ✗ | 0.866 | 0.019 | SB |
| News, Documentary | ✓ | < 0.001 | 1.469 | SB |
| Discussion, Science/Education | ✓ | < 0.001 | -0.485 | SB |
| Discussion, Documentary | ✓ | < 0.001 | 1.23 | SB |
| Science/Education, Documentary | ✓ | < 0.001 | 1.34 | SB |
| News, Discussion | ✓ | < 0.001 | -0.684 | SM |
| News, Science/Education | ✗ | 0.878 | 0.018 | SM |
| News, Documentary | ✓ | < 0.001 | -1.14 | SM |
| Discussion, Science/Education | ✓ | < 0.001 | 0.672 | SM |
| Discussion, Documentary | ✓ | < 0.001 | -0.535 | SM |
| Science/Education, Documentary | ✓ | < 0.001 | -1.105 | SM |
| News, Discussion | ✗ | 0.121 | 0.198 | SE |
| News, Science/Education | ✗ | 0.494 | -0.079 | SE |
| News, Documentary | ✓ | < 0.001 | -0.77 | SE |
| Discussion, Science/Education | ✓ | 0.002 | -0.273 | SE |
| Discussion, Documentary | ✓ | < 0.001 | -1.082 | SE |
| Science/Education, Documentary | ✓ | < 0.001 | -0.647 | SE |
| Dialog, Monolog | ✓ | < 0.001 | -0.473 | SB |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | 1.353 | SB |
| Monolog, Cooperative Monolog | ✓ | < 0.001 | 1.621 | SB |
| Dialog, Monolog | ✓ | < 0.001 | 0.541 | SM |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | -0.754 | SM |
| Monolog, Cooperative Monolog | ✓ | < 0.001 | -1.234 | SM |

Fortsetzung auf nächster Seite ...

| Data | < 0.05 | P-Value | EG | Statistic |
|------------------------------|--------|---------|--------|-----------|
| Dialog, Monolog | ✗ | 0.281 | -0.085 | SE |
| Dialog, Cooperative Monolog | ✓ | < 0.001 | -1.017 | SE |
| Monolog, Cooperative Monolog | ✓ | < 0.001 | -0.825 | SE |

Tabelle B.4: Zu Abschnitt 7.3: Vergleich der DM pro Position im Satz

| Data | < 0.05 | P-Value | EG | Statistic |
|---------------------------------|--------|---------|--------|-------------|
| Oral-Acoustic, Literate-Written | ✗ | 0.937 | 0.001 | Temporal |
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | 0.322 | Contingency |
| Oral-Acoustic, Literate-Written | ✓ | < 0.001 | 0.375 | Comparison |
| Oral-Acoustic, Literate-Written | ✗ | 0.432 | -0.009 | Expansion |
| News, Discussion | ✗ | 0.391 | 0.104 | Temporal |
| News, Science/Education | ✗ | 0.243 | -0.115 | Temporal |
| News, Documentary | ✗ | 0.124 | -0.111 | Temporal |
| News, Presentation | ✗ | 0.079 | -0.118 | Temporal |
| Discussion, Science/Education | ✗ | 0.034 | -0.183 | Temporal |
| Discussion, Documentary | ✓ | < 0.001 | -0.18 | Temporal |
| Discussion, Presentation | ✓ | < 0.001 | -0.179 | Temporal |
| Science/Education, Documentary | ✗ | 0.948 | 0.004 | Temporal |
| Science/Education, Presentation | ✗ | 0.853 | -0.011 | Temporal |
| Documentary, Presentation | ✗ | 0.632 | -0.015 | Temporal |
| News, Discussion | ✗ | 0.75 | -0.04 | Contingency |
| News, Science/Education | ✗ | 0.516 | -0.073 | Contingency |
| News, Documentary | ✗ | 0.465 | 0.08 | Contingency |
| News, Presentation | ✓ | < 0.001 | -0.336 | Contingency |
| Discussion, Science/Education | ✗ | 0.646 | -0.04 | Contingency |
| Discussion, Documentary | ✗ | 0.075 | 0.127 | Contingency |
| Discussion, Presentation | ✓ | < 0.001 | -0.316 | Contingency |
| Science/Education, Documentary | ✗ | 0.033 | 0.147 | Contingency |
| Science/Education, Presentation | ✓ | < 0.001 | -0.261 | Contingency |
| Documentary, Presentation | ✓ | < 0.001 | -0.402 | Contingency |
| News, Discussion | ✓ | < 0.001 | -0.65 | Comparison |
| News, Science/Education | ✓ | < 0.001 | -0.327 | Comparison |
| News, Documentary | ✓ | < 0.001 | -0.44 | Comparison |
| News, Presentation | ✗ | 0.746 | 0.035 | Comparison |
| Discussion, Science/Education | ✓ | < 0.001 | 0.426 | Comparison |
| Discussion, Documentary | ✓ | < 0.001 | 0.367 | Comparison |
| Discussion, Presentation | ✓ | < 0.001 | 0.665 | Comparison |
| Science/Education, Documentary | ✗ | 0.151 | -0.089 | Comparison |
| Science/Education, Presentation | ✓ | < 0.001 | 0.354 | Comparison |
| Documentary, Presentation | ✓ | < 0.001 | 0.466 | Comparison |
| News, Discussion | ✓ | < 0.001 | 0.547 | Expansion |
| News, Science/Education | ✗ | 0.181 | 0.149 | Expansion |
| News, Documentary | ✗ | 0.408 | 0.09 | Expansion |
| News, Presentation | ✓ | 0.001 | 0.446 | Expansion |
| Discussion, Science/Education | ✓ | 0.002 | -0.258 | Expansion |
| Discussion, Documentary | ✓ | < 0.001 | -0.471 | Expansion |
| Discussion, Presentation | ✓ | 0.005 | -0.161 | Expansion |

Fortsetzung auf nächster Seite ...

| Data | < 0.05 | P-Value | EG | Statistic |
|---------------------------------|--------|---------|--------|-------------|
| Science/Education, Documentary | ✗ | 0.279 | -0.075 | Expansion |
| Science/Education, Presentation | ✗ | 0.018 | 0.188 | Expansion |
| Documentary, Presentation | ✓ | < 0.001 | 0.363 | Expansion |
| Dialog, Monolog | ✗ | 0.646 | -0.036 | Temporal |
| Dialog, Cooperative-Monolog | ✗ | 0.051 | -0.111 | Temporal |
| Dialog, Speech | ✗ | 0.04 | -0.111 | Temporal |
| Monolog, Cooperative-Monolog | ✗ | 0.036 | -0.096 | Temporal |
| Monolog, Speech | ✗ | 0.029 | -0.098 | Temporal |
| Cooperative-Monolog, Speech | ✗ | 0.796 | -0.008 | Temporal |
| Dialog, Monolog | ✗ | 0.296 | -0.081 | Contingency |
| Dialog, Cooperative-Monolog | ✗ | 0.143 | 0.089 | Contingency |
| Dialog, Speech | ✓ | < 0.001 | -0.337 | Contingency |
| Monolog, Cooperative-Monolog | ✓ | 0.016 | 0.16 | Contingency |
| Monolog, Speech | ✓ | < 0.001 | -0.253 | Contingency |
| Cooperative-Monolog, Speech | ✓ | < 0.001 | -0.402 | Contingency |
| Dialog, Monolog | ✓ | < 0.001 | 0.502 | Comparison |
| Dialog, Cooperative-Monolog | ✓ | < 0.001 | 0.309 | Comparison |
| Dialog, Speech | ✓ | < 0.001 | 0.621 | Comparison |
| Monolog, Cooperative-Monolog | ✓ | < 0.001 | -0.268 | Comparison |
| Monolog, Speech | ✓ | 0.001 | 0.221 | Comparison |
| Cooperative-Monolog, Speech | ✓ | < 0.001 | 0.465 | Comparison |
| Dialog, Monolog | ✓ | < 0.001 | -0.281 | Expansion |
| Dialog, Cooperative-Monolog | ✓ | < 0.001 | -0.401 | Expansion |
| Dialog, Speech | ✗ | 0.409 | -0.047 | Expansion |
| Monolog, Cooperative-Monolog | ✗ | 0.823 | -0.015 | Expansion |
| Monolog, Speech | ✓ | < 0.001 | 0.258 | Expansion |
| Cooperative-Monolog, Speech | ✓ | < 0.001 | 0.866 | Expansion |

Tabelle B.5: Zu Abschnitt 7.4: Vergleich der DM-Klassen pro Textsorte

Anmerkung: Im Folgenden werden die in Abschnitt 7.5 referenzierten Tabellen und Diagramme aufgeführt. Die Tabellen zeigen für jeden der häufigsten DM der einzelnen Klassen die prozentuale Zu- oder Abnahme in der Nutzung von der erstgenannten Textsorte zur zweitgenannten.

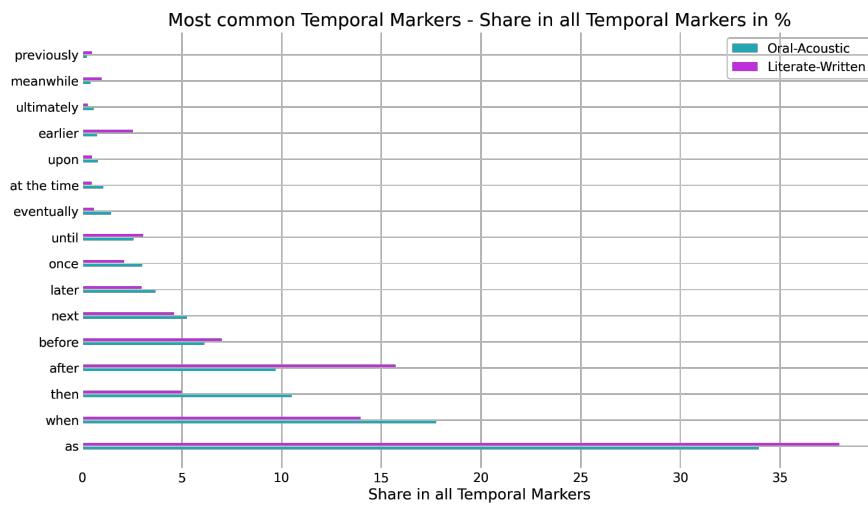


Abbildung B.1: Diskursarten - Häufigste TEMPORAL DM

| DM | OA-LS |
|-------------|--------|
| as | 11.9 |
| when | -21.4 |
| then | -52.37 |
| after | 62.21 |
| before | 14.22 |
| next | -12.1 |
| later | -18.93 |
| once | -29.84 |
| until | 18.97 |
| eventually | -59.49 |
| at the time | -54.22 |
| upon | -37.11 |
| earlier | 240.02 |
| ultimately | -49.64 |
| meanwhile | 131.57 |
| previously | 104.26 |

Tabelle B.6: Prozentuale Zu- und Abnahme der Benutzung einzelner TEMPORAL DM von der erstgenannten Textsorte zur zweitgenannten

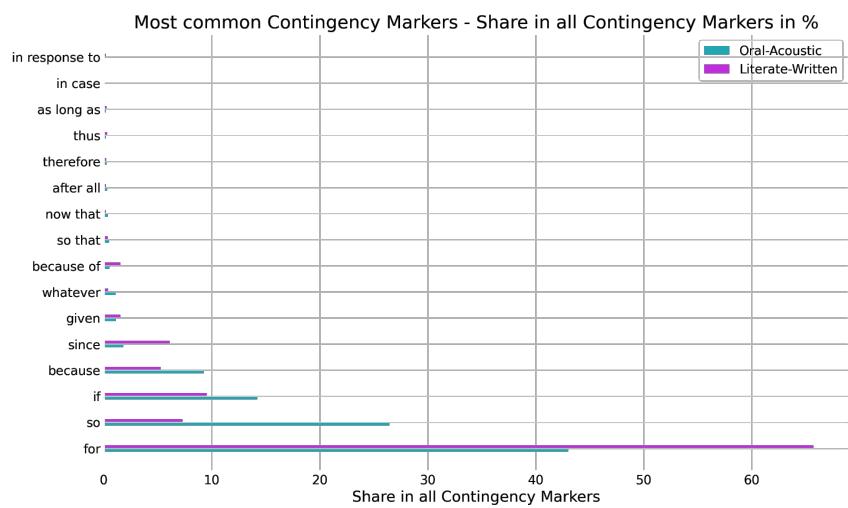


Abbildung B.2: Diskursarten - Häufigste CONTINGENCY DM

| DM | OA-LS |
|----------------|--------|
| for | 52.73 |
| so | -72.36 |
| if | -32.97 |
| because | -42.94 |
| since | 237.97 |
| given | 37.76 |
| whatever | -63.04 |
| because of | 195.46 |
| so that | -25.68 |
| now that | -49.17 |
| after all | -35.09 |
| therefore | -16.85 |
| thus | 43.51 |
| as long as | 33.25 |
| in case | -6.08 |
| in response to | 348.62 |

Tabelle B.7: Prozentuale Zu- und Abnahme der Benutzung einzelner CONTINGENCY DM von der erstgenannten Textsorte zur zweitgenannten

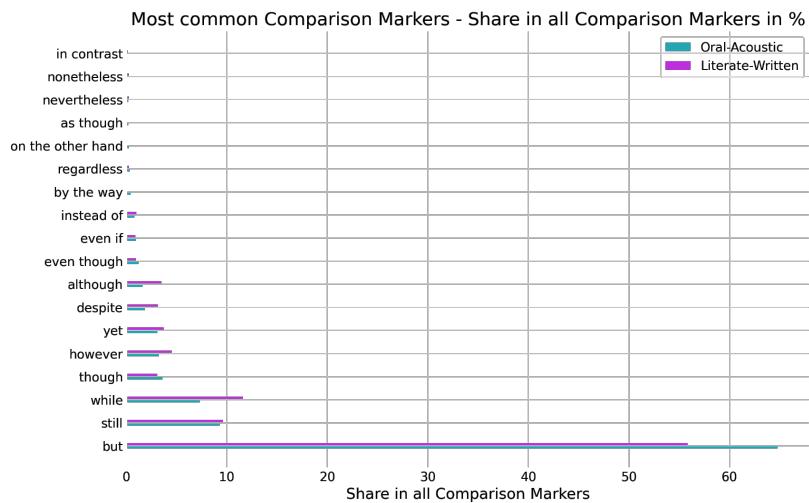


Abbildung B.8: Diskursarten - Häufigste COMPARISON DM

| DM | OA-LS |
|-------------------|--------|
| but | -13.82 |
| still | 3.43 |
| while | 58.79 |
| though | -14.54 |
| however | 39.68 |
| yet | 20.62 |
| despite | 67.77 |
| although | 118.06 |
| even though | -19.87 |
| even if | -5.09 |
| instead of | 21.54 |
| by the way | -81.62 |
| regardless | -27.51 |
| on the other hand | -55.14 |
| as though | -46.79 |
| nevertheless | 42.45 |
| nonetheless | 242.23 |
| in contrast | 288.33 |

Tabelle B.8: Prozentuale Zu- und Abnahme der Benutzung einzelner COMPARISON DM von der erstgenannten Textsorte zur zweitgenannten

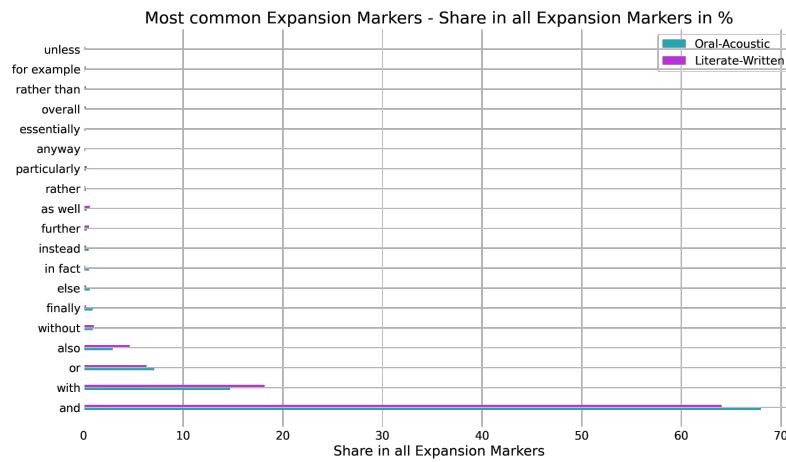


Abbildung B.4: Diskursarten - Häufigste EXPANSION DM

| DM | OA-LS |
|--------------|--------|
| and | -5.82 |
| with | 23.64 |
| or | 10.81 |
| also | 58.93 |
| without | 13.22 |
| finally | -70.4 |
| else | -57.95 |
| in fact | -68.14 |
| instead | -47.68 |
| further | 68.57 |
| as well | 96.5 |
| rather | -34.17 |
| particularly | 45.46 |
| anyway | -63.74 |
| essentially | -61.64 |
| overall | 274.34 |
| rather than | 87.78 |
| for example | 40.59 |
| unless | 99.34 |

Tabelle B.9: Prozentuale Zu- und Abnahme der Benutzung einzelner EXPANSION DM von der erstgenannten Textsorte zur zweitgenannten

Abkürzungen:

- Ne - News
- Do - Documentary
- Di - Discussion
- Pr - Presentation
- Sc - Science/Education

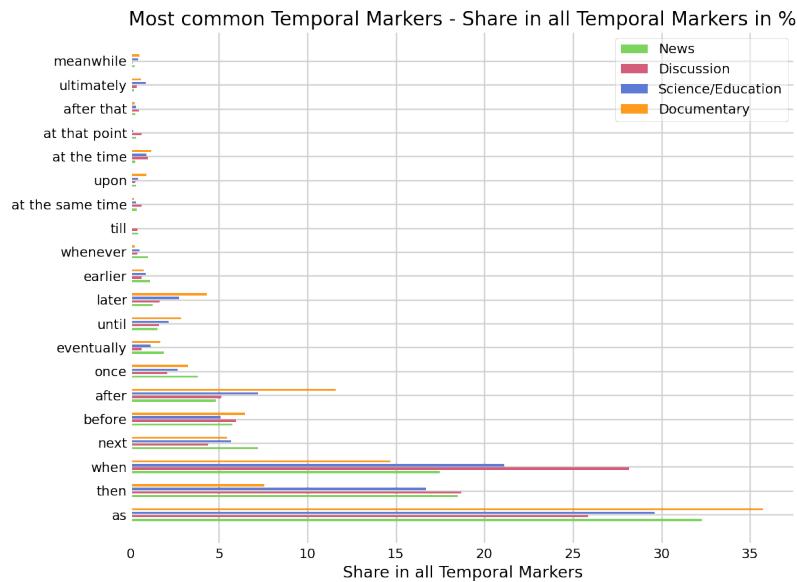


Abbildung B.5: Genres - Häufigste TEMPORAL DM

| DM | Ne-Di | Ne-Sc | Ne-Do | Ne-Pr | Di-Sc | Di-Do | Di-Pr | Sc-Do | Sc-Pr | Do-Pr |
|------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| as | -19.91 | -8.24 | 10.72 | -0.22 | 14.57 | 38.25 | 24.59 | 20.66 | 8.74 | -9.88 |
| then | 1.03 | -9.79 | -59.1 | -15.69 | -10.71 | -59.52 | -16.55 | -54.67 | -6.54 | 106.17 |
| when | 61.03 | 20.76 | -15.82 | 38.04 | -25.01 | -47.73 | -14.27 | -30.29 | 14.31 | 63.99 |
| next | -39.0 | -21.3 | -24.54 | -38.22 | 29.02 | 23.71 | 1.29 | -4.11 | -21.49 | -18.12 |
| before | 3.5 | -11.41 | 12.17 | -10.68 | -14.4 | 8.87 | -13.71 | 26.61 | 0.81 | -20.37 |
| after | 6.83 | 49.48 | 140.06 | 10.8 | 39.93 | 124.72 | 3.72 | 60.59 | -25.88 | -53.85 |
| once | -45.15 | -30.44 | -14.72 | -36.3 | 26.81 | 55.46 | 16.12 | 22.59 | -8.43 | -25.31 |
| eventually | -66.89 | -39.4 | -9.37 | -54.02 | 83.0 | 173.7 | 38.87 | 49.56 | -24.12 | -49.26 |
| until | 5.18 | 41.15 | 85.64 | 30.46 | 34.19 | 76.5 | 24.04 | 31.52 | -7.57 | -29.72 |
| later | 32.96 | 121.46 | 245.63 | 96.94 | 66.56 | 159.95 | 48.12 | 56.07 | -11.07 | -43.02 |
| earlier | -43.71 | -22.27 | -30.12 | -43.93 | 38.08 | 24.13 | -0.39 | -10.1 | -27.86 | -19.75 |
| whenever | -56.48 | -48.05 | -74.33 | -63.4 | 19.87 | -41.01 | -15.91 | -50.58 | -29.55 | 42.56 |
| till | -10.21 | -83.14 | -82.06 | -64.93 | -81.23 | -80.02 | -60.95 | 6.42 | 108.03 | 95.48 |
| at the same time | 86.44 | -8.32 | -37.94 | 119.18 | -50.83 | -66.71 | 17.56 | -32.31 | 139.07 | 253.16 |
| upon | -18.76 | 32.67 | 181.49 | 91.78 | 63.32 | 246.52 | 136.08 | 112.17 | 44.55 | -31.87 |
| at the time | 241.1 | 210.8 | 301.4 | 139.72 | -8.88 | 17.68 | -29.72 | 29.15 | -22.87 | -40.28 |
| at that point | 92.22 | -43.81 | -72.7 | -42.77 | -70.77 | -85.8 | -70.23 | -51.42 | 1.85 | 109.67 |
| after that | 70.25 | 14.25 | -12.32 | 5.2 | -32.89 | -48.5 | -38.21 | -23.26 | -7.92 | 19.98 |
| ultimately | 73.34 | 288.65 | 170.23 | 128.31 | 124.21 | 55.89 | 31.71 | -30.47 | -41.26 | -15.51 |
| meanwhile | -48.51 | 68.57 | 104.81 | -53.82 | 227.4 | 297.78 | -10.3 | 21.5 | -72.6 | -77.45 |

Tabelle B.10: Prozentuale Zu- und Abnahme der Benutzung einzelner TEMPORAL DM von der erstgenannten Textsorte zur zweitgenannten

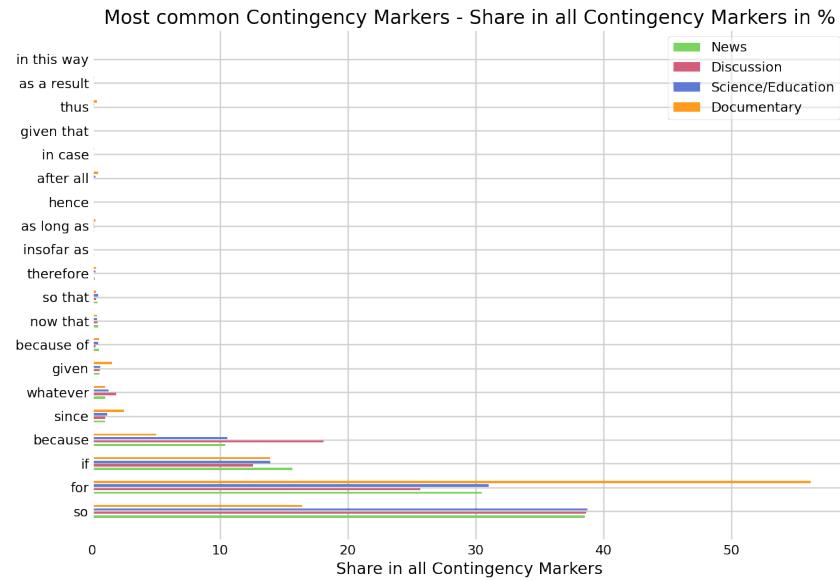


Abbildung B.6: Genres - Häufigste CONTINGENCY DM

| DM | Ne-Di | Ne-Sc | Ne-Do | Ne-Pr | Di-Sc | Di-Do | Di-Pr | Sc-Do | Sc-Pr | Do-Pr |
|-------------|--------|--------|---------|---------|--------|--------|--------|--------|--------|--------|
| so | 0.31 | 0.6 | -57.37 | -9.68 | 0.29 | -57.5 | -9.96 | -57.63 | -10.22 | 111.87 |
| for | -15.82 | 1.7 | 84.39 | 0.27 | 20.81 | 119.04 | 19.1 | 81.3 | -1.41 | -45.62 |
| if | -19.53 | -10.93 | -11.12 | 0.71 | 10.68 | 10.45 | 25.15 | -0.21 | 13.07 | 13.31 |
| because | 73.77 | 1.67 | -51.8 | 20.88 | -41.49 | -72.26 | -30.43 | -52.59 | 18.9 | 150.77 |
| since | -2.88 | 16.56 | 140.88 | 14.56 | 20.01 | 148.02 | 17.96 | 106.66 | -1.71 | -52.44 |
| whatever | 88.88 | 29.22 | 2.3 | -24.23 | -31.59 | -45.84 | -59.88 | -20.83 | -41.36 | -25.93 |
| given | -7.19 | 9.12 | 153.6 | 32.85 | 17.58 | 173.26 | 43.15 | 132.4 | 21.75 | -47.61 |
| because of | -45.68 | -11.1 | 4.92 | 39.95 | 63.65 | 93.14 | 157.63 | 18.02 | 57.42 | 33.39 |
| now that | -12.84 | -21.91 | -15.96 | -10.87 | -10.41 | -3.58 | 2.26 | 7.63 | 14.14 | 6.06 |
| so that | -20.58 | 26.69 | -14.29 | 165.79 | 59.53 | 7.92 | 234.67 | -32.35 | 109.79 | 210.11 |
| therefore | -61.45 | 23.41 | 33.88 | 57.67 | 220.15 | 247.28 | 309.01 | 8.48 | 27.76 | 17.77 |
| insofar as | -98.82 | -96.39 | -96.12 | -97.82 | 205.92 | 228.8 | 85.13 | 7.48 | -39.48 | -43.7 |
| as long as | 13.56 | 14.27 | 98.56 | 11.57 | 0.63 | 74.86 | -1.75 | 73.76 | -2.37 | -43.81 |
| hence | -95.2 | -75.51 | -57.45 | -68.39 | 409.86 | 785.94 | 558.24 | 73.76 | 29.1 | -25.7 |
| after all | 21.03 | 462.22 | 1028.73 | 298.33 | 364.54 | 832.62 | 229.12 | 100.76 | -29.15 | -64.71 |
| in case | 176.42 | 208.54 | 485.45 | 259.6 | 11.62 | 111.8 | 30.09 | 89.75 | 16.55 | -38.58 |
| given that | -41.73 | 16.56 | 76.86 | 2.35 | 100.02 | 203.51 | 75.64 | 51.74 | -12.19 | -42.13 |
| thus | -59.66 | 270.24 | 742.55 | 57.67 | 817.75 | 1988.5 | 290.83 | 127.57 | -57.41 | -81.29 |
| as a result | 101.71 | 722.76 | 1705.47 | 1213.92 | 307.89 | 795.07 | 551.38 | 119.44 | 59.7 | -27.23 |
| in this way | -25.29 | -20.01 | 110.84 | 268.82 | 7.07 | 182.22 | 393.68 | 163.58 | 361.08 | 74.93 |

Tabelle B.11: Prozentuale Zu- und Abnahme der Benutzung einzelner CONTINGENCY DM von der erstgenannten Textsorte zur zweitgenannten

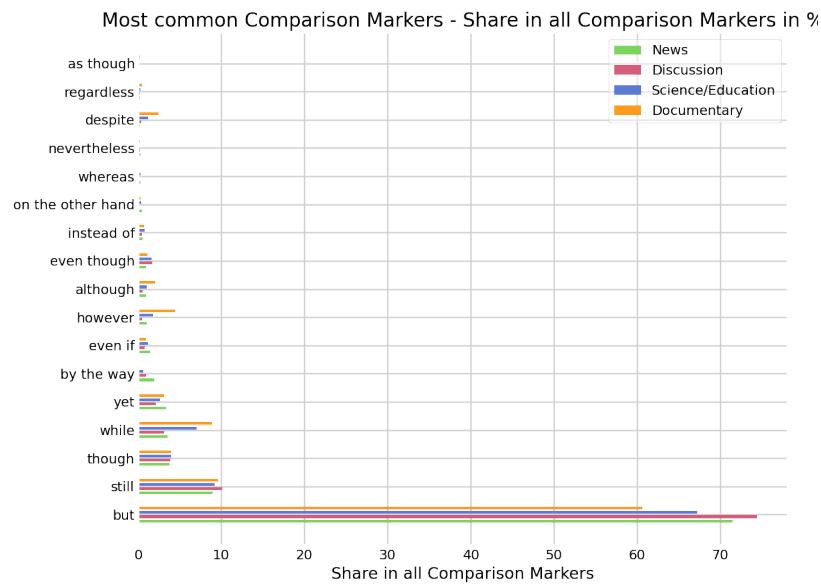


Abbildung B.7: Genres - Häufigste COMPARISON DM

| DM | Ne-Di | Ne-Sc | Ne-Do | Ne-Pr | Di-Sc | Di-Do | Di-Pr | Sc-Do | Sc-Pr | Do-Pr |
|-------------------|--------|--------|---------|--------|--------|--------|--------|--------|--------|--------|
| but | 4.05 | -5.98 | -15.17 | 2.65 | -9.64 | -18.47 | -1.35 | -9.78 | 9.18 | 21.01 |
| still | 13.61 | 2.96 | 8.07 | -17.63 | -9.38 | -4.88 | -27.5 | 4.96 | -20.0 | -23.78 |
| though | 3.41 | 4.74 | 5.61 | -53.03 | 1.29 | 2.13 | -54.58 | 0.83 | -55.16 | -55.53 |
| while | -10.37 | 99.31 | 152.19 | 15.88 | 122.37 | 181.37 | 29.29 | 26.53 | -41.86 | -54.05 |
| yet | -36.74 | -22.98 | -8.94 | 10.23 | 21.76 | 43.95 | 74.25 | 18.23 | 43.12 | 21.05 |
| by the way | -50.32 | -68.48 | -94.31 | -41.43 | -36.56 | -88.54 | 17.88 | -81.94 | 85.81 | 928.78 |
| even if | -42.25 | -17.96 | -34.41 | -32.8 | 42.05 | 13.56 | 16.35 | -20.05 | -18.09 | 2.46 |
| however | -57.86 | 71.84 | 322.22 | -2.14 | 307.74 | 901.85 | 132.19 | 145.71 | -43.05 | -76.82 |
| although | -44.8 | 0.59 | 101.86 | -7.25 | 82.25 | 265.71 | 68.05 | 100.67 | -7.79 | -54.05 |
| even though | 84.32 | 71.44 | 17.37 | 31.17 | -6.99 | -36.32 | -28.84 | -31.54 | -23.49 | 11.75 |
| instead of | -21.27 | 49.51 | 35.3 | 161.2 | 89.91 | 71.85 | 231.78 | -9.51 | 74.7 | 93.06 |
| on the other hand | -80.97 | -18.89 | -40.05 | -6.31 | 328.78 | 214.96 | 392.26 | -26.55 | 14.8 | 56.29 |
| whereas | -60.47 | -14.02 | -65.81 | 20.12 | 117.5 | -13.51 | 203.86 | -60.23 | 39.71 | 251.31 |
| nevertheless | -30.79 | -56.45 | -16.12 | -61.89 | -37.08 | 21.2 | -44.21 | 92.62 | -11.33 | -53.97 |
| despite | 125.68 | 597.99 | 1323.22 | 392.55 | 209.29 | 530.65 | 118.26 | 103.9 | -29.43 | -65.39 |
| regardless | 22.35 | 59.67 | 143.92 | 40.98 | 30.5 | 99.35 | 15.23 | 52.76 | -11.7 | -42.2 |
| as though | -17.52 | 70.32 | 228.0 | 145.68 | 106.51 | 297.69 | 197.88 | 92.58 | 44.25 | -25.1 |

Tabelle B.12: Prozentuale Zu- und Abnahme der Benutzung einzelner COMPARISON DM von der erstgenannten Textsorte zur zweitgenannten

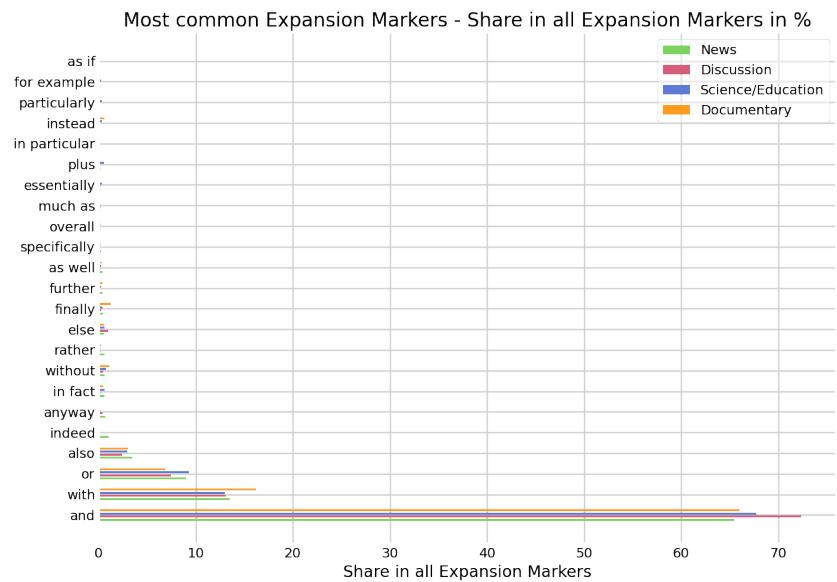


Abbildung B.8: Genres - Häufigste EXPANSION DM

| DM | Ne-Di | Ne-Sc | Ne-Do | Ne-Pr | Di-Sc | Di-Do | Di-Pr | Sc-Do | Sc-Pr | Do-Pr |
|---------------|--------|--------|---------|--------|--------|--------|--------|--------|--------|--------|
| and | 10.56 | 3.5 | 0.9 | 9.76 | -6.38 | -8.74 | -0.72 | -2.52 | 6.05 | 8.78 |
| with | -3.36 | -3.96 | 19.81 | -11.42 | -0.62 | 23.97 | -8.35 | 24.75 | -7.77 | -26.07 |
| or | -17.39 | 2.79 | -24.16 | -25.99 | 24.43 | -8.19 | -10.41 | -26.22 | -28.0 | -2.42 |
| also | -31.65 | -15.07 | -13.2 | -20.68 | 24.26 | 27.0 | 16.05 | 2.21 | -6.61 | -8.63 |
| indeed | -96.72 | -93.72 | -88.62 | -81.95 | 91.17 | 246.38 | 449.38 | 81.19 | 187.38 | 58.61 |
| anyway | -42.03 | -67.03 | -80.71 | -76.46 | -43.13 | -66.73 | -59.39 | -41.51 | -28.6 | 22.06 |
| in fact | -67.76 | -4.21 | -21.26 | 51.35 | 197.11 | 144.22 | 369.44 | -17.8 | 58.0 | 92.22 |
| without | -21.9 | 22.28 | 77.02 | 23.15 | 56.57 | 126.67 | 57.68 | 44.77 | 0.71 | -30.48 |
| rather | -63.27 | -68.11 | -60.38 | -48.8 | -13.19 | 7.87 | 39.37 | 24.25 | 60.54 | 29.2 |
| else | 65.08 | 11.97 | -4.39 | 8.0 | -32.17 | -42.08 | -34.58 | -14.61 | -3.55 | 12.95 |
| finally | -34.7 | -6.6 | 177.03 | -13.82 | 43.03 | 324.24 | 31.97 | 196.6 | -7.74 | -68.89 |
| further | -78.98 | -29.02 | 7.42 | -48.83 | 237.71 | 411.07 | 143.46 | 51.33 | -27.91 | -52.36 |
| as well | -53.8 | -27.65 | -17.49 | -20.78 | 56.6 | 78.58 | 71.45 | 14.03 | 9.48 | -3.99 |
| specifically | -62.85 | -27.15 | -41.37 | -66.05 | 96.07 | 57.81 | -8.63 | -19.51 | -53.4 | -42.1 |
| overall | -73.67 | -20.57 | -77.7 | -73.67 | 201.65 | -15.31 | 0.0 | -71.93 | -66.85 | 18.08 |
| much as | 1.59 | -50.58 | -34.73 | -39.57 | -51.35 | -35.75 | -40.52 | 32.07 | 22.27 | -7.42 |
| essentially | -29.78 | 33.82 | -39.6 | -16.81 | 90.57 | -13.99 | 18.47 | -54.87 | -37.83 | 37.75 |
| plus | -10.48 | 245.43 | -35.09 | -55.07 | 285.86 | -27.49 | -49.81 | -81.21 | -86.99 | -30.78 |
| in particular | 68.73 | 18.98 | -4.84 | 22.25 | -29.49 | -43.6 | -27.55 | -20.02 | 2.75 | 28.46 |
| instead | 108.7 | 536.97 | 1101.52 | 521.27 | 205.21 | 475.72 | 197.69 | 88.63 | -2.47 | -48.29 |
| particularly | 57.0 | 302.23 | 197.84 | 191.31 | 156.2 | 89.71 | 85.55 | -25.95 | -27.58 | -2.19 |
| for example | -56.34 | 75.4 | -42.49 | 77.03 | 301.7 | 31.71 | 305.43 | -67.21 | 0.93 | 207.82 |
| as if | 45.17 | 6.52 | 121.56 | 87.3 | -26.62 | 52.63 | 29.03 | 108.01 | 75.85 | -15.46 |

Tabelle B.13: Prozentuale Zu- und Abnahme der Benutzung einzelner EXPANSION DM von der erstgenannten Textsorte zur zweitgenannten

- **Di** - Dialog
- **Mo** - Monolog
- **Co** - Cooperative Monolog
- **Sp** - Speech

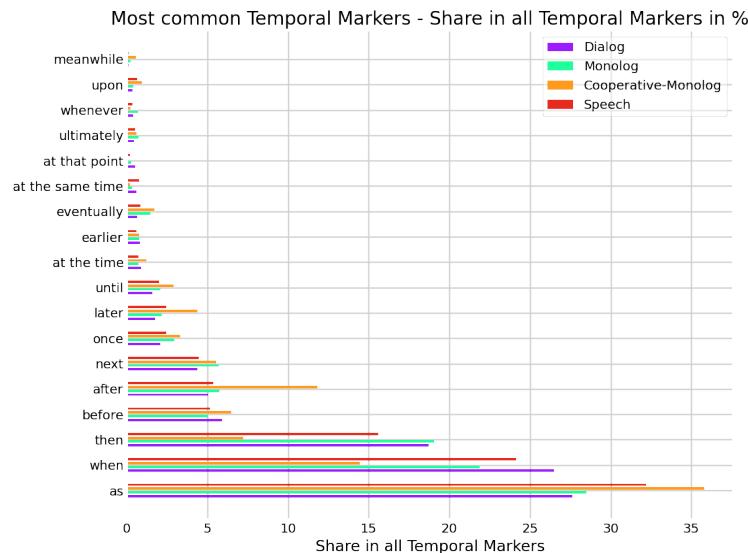


Abbildung B.9: Konversationsarten - Häufigste TEMPORAL DM

| DM | Di-Mo | Di-Co | Di-Sp | Mo-Co | Mo-Sp | Co-Sp |
|------------------|--------|--------|--------|--------|--------|--------|
| as | 2.99 | 29.54 | 16.45 | 25.78 | 13.08 | -10.1 |
| when | -17.38 | -45.39 | -8.88 | -33.9 | 10.29 | 66.85 |
| then | 1.81 | -61.36 | -16.64 | -62.05 | -18.12 | 115.75 |
| before | -14.71 | 9.4 | -12.98 | 28.27 | 2.03 | -20.46 |
| after | 14.5 | 135.12 | 6.51 | 105.35 | -6.98 | -54.7 |
| next | 30.18 | 26.22 | 1.62 | -3.04 | -21.94 | -19.49 |
| once | 42.84 | 59.71 | 17.81 | 11.81 | -17.53 | -26.24 |
| later | 24.23 | 153.27 | 42.16 | 103.86 | 14.43 | -43.87 |
| until | 31.6 | 81.82 | 26.55 | 38.16 | -3.84 | -30.4 |
| at the time | -16.59 | 35.05 | -20.49 | 61.91 | -4.68 | -41.13 |
| earlier | -2.66 | -2.95 | -21.61 | -0.29 | -19.47 | -19.23 |
| eventually | 121.68 | 161.95 | 31.71 | 18.17 | -40.59 | -49.72 |
| at the same time | -46.05 | -65.62 | 26.1 | -36.26 | 133.74 | 266.74 |
| at that point | -49.81 | -83.98 | -64.19 | -68.08 | -28.65 | 123.54 |
| ultimately | 60.96 | 35.74 | 13.94 | -15.67 | -29.21 | -16.06 |
| whenever | 71.43 | -38.93 | -12.7 | -64.37 | -49.07 | 42.94 |
| upon | 17.2 | 168.35 | 82.93 | 128.97 | 56.09 | -31.83 |
| meanwhile | 112.19 | 410.89 | 8.94 | 140.77 | -48.66 | -78.68 |

Tabelle B.14: Prozentuale Zu- und Abnahme der Benutzung einzelner TEMPORAL DM von der erstgenannten Textsorte zur zweitgenannten

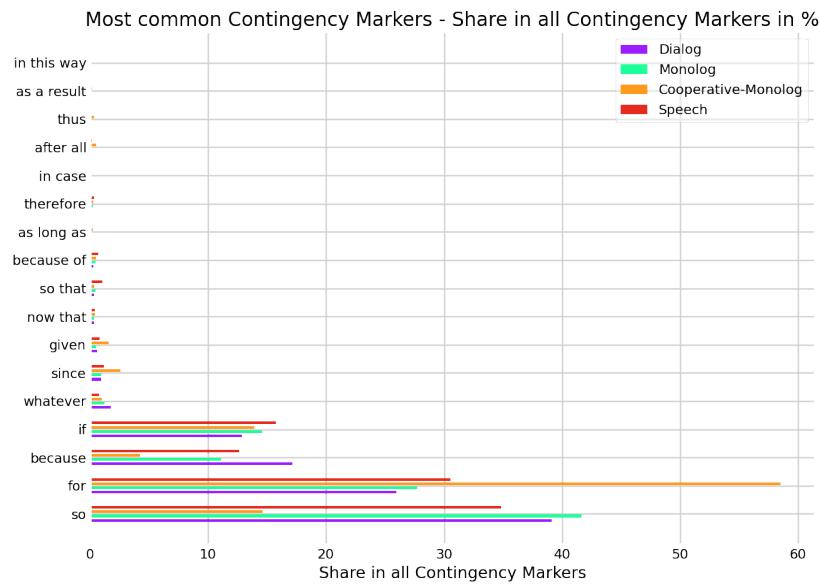


Abbildung B.10: Konversationsarten - Häufigste CONTINGENCY DM

| DM | Di-Mo | Di-Co | Di-Sp | Mo-Co | Mo-Sp | Co-Sp |
|-------------|--------|---------|--------|--------|--------|--------|
| so | 6.55 | -62.62 | -10.94 | -64.92 | -16.41 | 138.25 |
| for | 6.7 | 125.3 | 17.73 | 111.15 | 10.34 | -47.74 |
| because | -35.46 | -75.3 | -26.55 | -61.73 | 13.79 | 197.34 |
| if | 13.17 | 8.4 | 22.6 | -4.21 | 8.33 | 13.09 |
| whatever | -30.02 | -44.43 | -57.24 | -20.59 | -38.89 | -23.05 |
| since | 2.85 | 173.36 | 23.43 | 165.77 | 20.01 | -54.85 |
| given | -11.51 | 172.25 | 36.71 | 207.67 | 54.5 | -49.79 |
| now that | -12.17 | 5.25 | 7.32 | 19.84 | 22.19 | 1.97 |
| so that | 46.08 | 0.31 | 212.71 | -31.34 | 114.07 | 211.76 |
| because of | 72.78 | 85.02 | 142.81 | 7.08 | 40.53 | 31.24 |
| as long as | -33.4 | 88.31 | -0.13 | 182.73 | 49.95 | -46.96 |
| therefore | 102.14 | 141.58 | 179.01 | 19.51 | 38.03 | 15.49 |
| in case | -8.09 | 119.52 | 31.34 | 138.83 | 42.89 | -40.17 |
| after all | 128.36 | 780.13 | 188.66 | 285.42 | 26.41 | -67.2 |
| thus | 163.39 | 1152.61 | 118.13 | 375.57 | -17.18 | -82.59 |
| as a result | 73.17 | 684.0 | 433.21 | 352.73 | 207.91 | -31.99 |
| in this way | 38.33 | 204.55 | 418.1 | 120.16 | 274.54 | 70.12 |

Tabelle B.15: Prozentuale Zu- und Abnahme der Benutzung einzelner CONTINGENCY DM von der erstgenannten Textsorte zur zweitgenannten

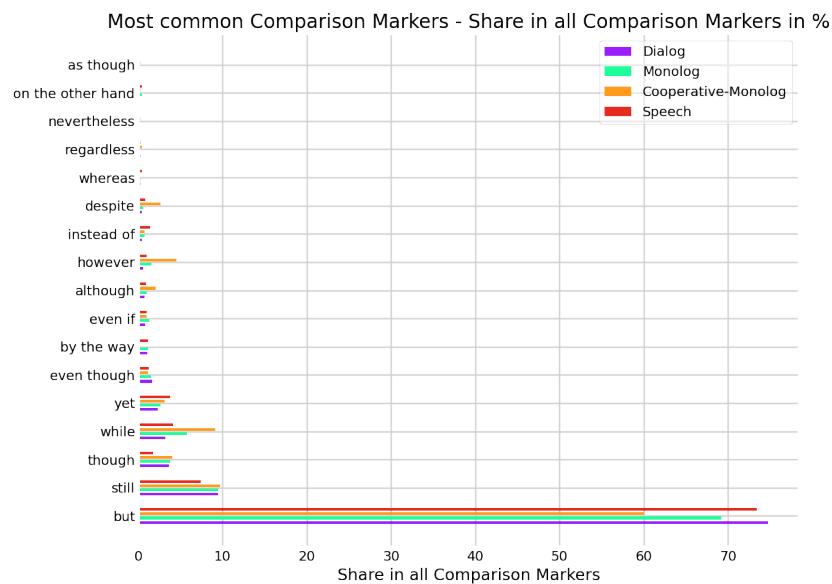


Abbildung B.11: Konversationsarten - Häufigste COMPARISON DM

| DM | Di-Mo | Di-Co | Di-Sp | Mo-Co | Mo-Sp | Co-Sp |
|-------------------|--------|--------|--------|--------|--------|---------|
| but | -7.35 | -19.66 | -1.73 | -13.29 | 6.06 | 22.32 |
| still | 0.26 | 2.79 | -22.23 | 2.52 | -22.44 | -24.34 |
| though | 4.99 | 12.33 | -50.97 | 6.99 | -53.3 | -56.36 |
| while | 77.65 | 182.84 | 26.4 | 59.21 | -28.85 | -55.31 |
| yet | 13.39 | 33.62 | 60.78 | 17.84 | 41.79 | 20.33 |
| even though | -10.33 | -31.35 | -23.55 | -23.45 | -14.74 | 11.37 |
| by the way | 5.95 | -95.31 | 5.81 | -95.57 | -0.14 | 2156.16 |
| even if | 53.73 | 13.5 | 14.87 | -26.17 | -25.28 | 1.21 |
| although | 37.89 | 188.59 | 31.65 | 109.29 | -4.52 | -54.38 |
| however | 184.02 | 708.05 | 83.21 | 184.51 | -35.49 | -77.33 |
| instead of | 62.5 | 73.08 | 223.28 | 6.51 | 100.18 | 87.94 |
| despite | 52.59 | 584.57 | 127.7 | 348.64 | 49.23 | -66.74 |
| whereas | 1.85 | -66.88 | 46.55 | -67.48 | 43.89 | 342.46 |
| regardless | -22.23 | 105.96 | 15.39 | 164.83 | 48.37 | -43.97 |
| nevertheless | -40.55 | 67.07 | -27.87 | 181.02 | 21.33 | -56.83 |
| on the other hand | 246.34 | 158.69 | 288.01 | -23.31 | 12.03 | 49.99 |
| as though | 87.27 | 254.18 | 162.84 | 89.13 | 40.36 | -25.79 |

Tabelle B.16: Prozentuale Zu- und Abnahme der Benutzung einzelner COMPARISON DM von der erstgenannten Textsorte zur zweitgenannten

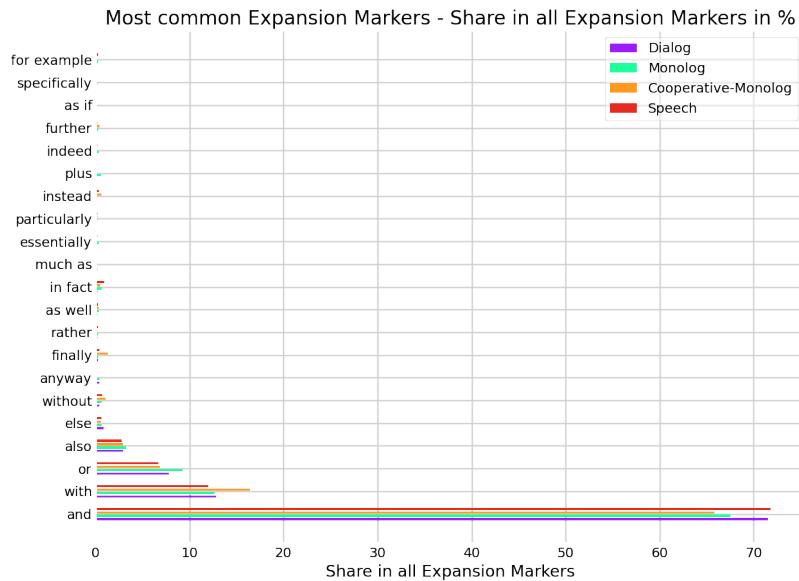


Abbildung B.12: Konversationsarten - Häufigste EXPANSION DM

| DM | Di-Mo | Di-Co | Di-Sp | Mo-Co | Mo-Sp | Co-Sp |
|--------------|--------|--------|--------|--------|--------|--------|
| and | -5.59 | -8.01 | 0.41 | -2.56 | 6.36 | 9.15 |
| with | -1.68 | 27.66 | -6.74 | 29.84 | -5.14 | -26.94 |
| or | 18.53 | -12.52 | -14.3 | -26.2 | -27.7 | -2.04 |
| also | 10.23 | -0.57 | -5.93 | -9.8 | -14.66 | -5.39 |
| else | -22.43 | -35.74 | -26.91 | -17.16 | -5.78 | 13.74 |
| without | 47.73 | 141.4 | 62.96 | 63.41 | 10.31 | -32.49 |
| anyway | -6.31 | -74.5 | -62.07 | -72.78 | -59.52 | 48.73 |
| finally | 5.53 | 356.52 | 36.75 | 332.59 | 29.58 | -70.04 |
| rather | 41.56 | 16.77 | 47.63 | -17.51 | 4.29 | 26.42 |
| as well | 65.15 | 63.18 | 55.62 | -1.19 | -5.77 | -4.63 |
| in fact | 233.38 | 148.54 | 364.02 | -25.45 | 39.19 | 86.7 |
| much as | -47.15 | -33.8 | -38.2 | 25.25 | 16.94 | -6.64 |
| essentially | 109.09 | -15.4 | 17.84 | -59.54 | -43.64 | 39.29 |
| particularly | 28.99 | 38.63 | 34.34 | 7.47 | 4.15 | -3.09 |
| instead | -4.5 | 334.19 | 115.95 | 354.65 | 126.13 | -50.26 |
| plus | 372.87 | -13.85 | -41.58 | -81.78 | -87.65 | -32.19 |
| indeed | 631.82 | 133.74 | 270.91 | -68.06 | -49.32 | 58.68 |
| further | 191.4 | 337.97 | 101.22 | 50.3 | -30.95 | -54.06 |
| as if | -32.56 | 69.98 | 40.55 | 152.03 | 108.4 | -17.31 |
| specifically | 61.78 | 31.33 | -25.25 | -18.82 | -53.79 | -43.08 |
| for example | 144.28 | -11.0 | 173.88 | -63.57 | 12.12 | 207.72 |

Tabelle B.17: Prozentuale Zu- und Abnahme der Benutzung einzelner EXPANSION DM von der erstgenannten Textsorte zur zweitgenannten

Literaturverzeichnis

Praneeth Bedapudi. Deepcorrection 1: Sentence segmentation of unpunctuated text., 2018. URL <https://praneethbedapudi.medium.com/deepcorrection-1-sentence-segmentation-of-unpunctuated-text-a1dbc0db4e98>. [Accessed: 16.12.2020]. 5.1.2.1, 5.1.2.1

Doug Beeferman, William Brannon, and Deb Roy. Radiotalk: a large-scale corpus of talk radio transcripts. *CoRR*, abs/1907.07073, 2019. URL <http://arxiv.org/abs/1907.07073>. 5.1.3.1, 5.1.4

Or Biran and Owen Rambow. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05, 12 2011. doi: 10.1142/S17938351X11001328. 3.1

Hardarik Blühdorn, Ad Foolen, and Óscar Loureda. Diskursmarker: Begriffs geschichte – theorie – beschreibung. ein bibliographischer Überblick. Diskursmarker im Deutschen. Reflexionen und Analysen, pages 4 – 47. Verlag für Gesprächsforschung, Göttingen, 2017. ISBN 978-3-936656-69-5. URL <http://nbn-resolving.de/urn:nbn:de:bsz:mh39-62874>. 3, 3.1, 3.1

Jeanne S. Chall and Harold E. Dial. Predicting listener understanding and interest in newscasts. *Educational Research Bulletin*, 27(6):141–168, 1948. ISSN 15554023. URL <http://www.jstor.org/stable/1473082>. 2.2

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online), December 2020. International

Committee on Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.coling-main.519>. 5.1.3.2, 5.1.3.2, 5.1.4

Debopam Das, Maite Taboada, and Paul McFetridge. Rst signalling corpus ldc2015t10, 2015. URL <https://catalog.ldc.upenn.edu/LDC2015T10.3.1>

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5042. URL <https://www.aclweb.org/anthology/W18-5042>. 3.1, 3.2, 3.2

Carl Jón Denbow. Listenability and readability: An experimental investigation. *Journalism Quarterly*, 52(2):285–290, 1975. URL <https://doi.org/10.1177/107769907505200213>. 2.2

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>. 5.1.2.1

William H. Dubay. The listenability of consumer- information phone scripts, 2007. 2.2

B. Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, C. Munteanu, A. Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare R. Voss, and F. Zeller. Automatic human utility evaluation of asr systems: does wer really predict performance? In *INTERSPEECH*, 2013. 5.1.2.2

Rudolf Flesch. *Marks of Readable Style: A Study in Adult Education*. Contributions to education. Columbia University, 1943. URL <https://books.google.de/books?id=ZQX0zQEACAAJ>. 2.2

Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3): 221–233, 1948. 2.2

Bruce Fraser. Types of english discourse markers. *Acta Linguistica Hungarica*, 38 (1/4):19–33, 1988. ISSN 12168076, 15882624. URL <http://www.jstor.org/stable/44362602>. 3.1

Bruce Fraser. An approach to discourse markers. *Journal of Pragmatics*, 14(3): 383 – 398, 1990. ISSN 0378-2166. URL <http://www.sciencedirect.com/science/article/pii/037821669090096V>. Special Issue: 'Selected papers from The International Pragmatics Conference, Antwerp, 17-22 August, 1987'. 3.1

Bruce Fraser. What are discourse markers? *Journal of Pragmatics*, 31(7):931 – 952, 1999. ISSN 0378-2166. URL <http://www.sciencedirect.com/science/article/pii/S0378216698001015>. Pragmatics: The Loaded Discipline? 3.1

Bruce Fraser. An account of discourse markers. *International Review of Pragmatics*, 1:293–320, 11 2009. 3, 3.1, 3.1, 3.1, 3.2, 3.2, 3.2, 8, 9

Kenneth A. Harwood. I. listenability and readability. *Speech Monographs*, 22(1): 49–53, 1955. URL <https://doi.org/10.1080/03637755509375133>. 2.2

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699, 2018. URL <http://arxiv.org/abs/1805.04699>. 5.1.3.3, 5.1.4

Jules Horne. Writing for audiobooks: Audio-first for flow and impact, 2019a. 1

Jules Horne. Writing for the ear - 10 tips from radio writing, 2019b. 2.3

Jules Horne. Writing for audiobooks - download, 2019c. 2.3, 3.1, 3.2

Katsunori Kotani and Takehiko Yoshimi. A listenability index consisting of subjective judgment and objective evaluation. In *Proceedings of the 2017 9th International Conference on Education Technology and Computers*, ICETC 2017, page 68–67, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450354356. URL <https://doi.org/10.1145/3175536.3175541>. 2.2

Katsunori Kotani, Shota Ueda, Takehiko Yoshimi, and Hiroaki Nanjo. A listenability measuring method for an adaptive computer-assisted language learning and teaching system. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 387–394, Phuket, Thailand, December 2014. Department of Linguistics, Chulalongkorn University. URL <https://www.aclweb.org/anthology/Y14-1045>. 2.2

Ian McCormick. *The Art of Connection: The Social Life of Sentences*. Quibble Academic, 2013. ISBN 978-1-49374841-9. 1, 2.3, 3.1, 3.2, 3.2

Josef Messerklinger. Listenability, 11 2006. 2.2

Katrin Ortmann and Stefanie Dipper. Variation between different discourse types: Literate vs. oral. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–79, Ann Arbor, Michigan, June 2019. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W19-1407>. 1, 2.1, 2.1, 2, 2.2, 2.3, 3.1, 4.1

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword fifth edition ldc2011t07, 2011. URL <https://catalog.ldc.upenn.edu/LDC2011T07>. 5.2.1, 5.2.2.1

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesel. The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011. 5.1.3.1, 5.1.3.3

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf. 3.1

Evan Sandhaus. The new york times annotated corpus ldc2008t19, 2008. URL <https://catalog.ldc.upenn.edu/LDC2008T19>. 5.2.1, 5.2.2.2

Deborah Schiffrin. *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press, 1987. doi: 10.1017/CBO9780511611841. 3.1

Wolf Schneider. *Deutsch für junge Profis - Wie man gut und lebendig schreibt*. Rowohlt Taschenbuch Verlag, 5 edition, 5 2011. ISBN 978-3-499-62629-6. 4.3, 7.1.2, 7.1.3

Wolf Schneider and Paul-Josef Raue. *Das neue Handbuch des Journalismus und des Online-Journalismus*. Rowohlt Taschenbuch Verlag, 1 edition, January 2012. ISBN 978-3-49962825-2. 1, 2.3

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5384–5396, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.coling-main.470>. 6.3.2.1

James R. Young. Understanding radio news: The effect of style. *Journalism Quarterly*, 27(1):19–23, 1950. URL <https://doi.org/10.1177/107769905002700103>. 2.2