

# **Diskursmarker**

in  
**schriftlichem & akustischem**  
**Diskurs**

BACHELORVERTEIDIGUNG

Johanna Sacher, 4.2.2021



- Es wird immer mehr gehört
- Unser Medienkonsum entwickelt sich weg vom Lesen, hin zum Hören
- Streamingdienste mit Hörbüchern und Podcast, Plattformen wie YouTube und auch Sprachassistenten machen es möglich
- Jeder kann heutzutage ganz einfach einen eigenen Podcast erstellen, ein eigenes Video hochladen
- mit audible kann man Bücher jederzeit von der Stelle an weiterhören, an der man mit Lesen aufgehört hat
- und Sprachassistenten lesen auf Wunsch die Zeitung vor
- Für Konsumenten ist das praktisch
- Doch für Produzenten und Forscher stellt diese Veränderung ein Problem dar:
- Die meisten unserer Tools, z.B. Analyseprogramme und Schreibassistenzsoftware, wurden für schriftliche Texte entwickelt, nicht für akustische
- schon intuitiv ist es so, dass wir anders reden, als wir schreiben, wir hören anders, als wir lesen
- Dieser Annahme gehe ich in meiner Arbeit auf den Grund
- Konkret habe ich am Beispiel von Diskursmarkern untersucht, wie sich akustischer und schriftlicher Diskurs unterscheiden
- Und meine Ergebnisse möchte ich Ihnen jetzt gerne vorstellen

Diese Arbeit liefert Evidenz für Unterschiede zwischen

- oral-akustischem und literat-schriftlichem Diskurs
- geskripteten und improvisierten oral-akustischen Texten
- interaktiven und passiven oral-akustischen Texten

Im Folgenden wird die unterschiedliche Verwendung von [Diskurmarkern](#)  
in den genannten Textsorten nachgewiesen

- Ich konnte feststellen dass diese intuitiv wahrgenommenen Unterschiede tatsächlich existieren,
- und Unterschiede zwischen oral-akustischem und literat-schriftlichem Diskurs nachweisen
- Des Weiteren habe ich auch Unterschiede zwischen geskripteten und improvisierten Texten gefunden, sowie zwischen interaktiven Texten wie z.B. Dialogen und eher passiven Texten, wie Monologen.
- Diese Ergebnisse basieren auf der Betrachtung von Diskurmarkern in den einzelnen Texten
- weil Diskurmarker einerseits von der relevanten Literatur in diesem Zusammenhang immer wieder genannt wurden und es andererseits aber bisher nicht besonders viel Forschung auf diesem Gebiet gibt

- oral-akustische Texte nutzen mehr Diskursmarker als literat-schriftliche
- improvisierte Texte nutzen mehr Diskursmarker als geskriptete
- interaktive Texte nutzen mehr Diskursmarker als passive

- Konkret konnte ich feststellen, dass oral-akustische Texte mehr Diskursmarker nutzen als literat-schriftliche,  
- dass improvisierte Texte mehr nutzen als geskriptete  
- und dass interaktive Texte mehr nutzen als passive

# BEGRIFF

## Diskursmarker

Wörter wie *and*, *but* und *so*

- keine inhaltliche Bedeutung
- signalisieren Beziehungen zwischen Diskurssegmenten
- Wegweiser im Text

Zuerst möchte ich kurz auf die relevanten Begriffe eingehen

- ganz häufig ist grade der Begriff „Diskursmarker“ gefallen
  - später genauer
  - für den Anfang reicht es zu wissen, dass Diskursmarker Wörter sind, die
    - keine inhaltliche Bedeutung zum Satz beitragen, sondern
    - als Wegweiser im Text dienen
    - indem sie Beziehungen zwischen Diskurssegmenten anzeigen

# BEGRIFF

## Diskurs

Quelle

Einheit von Sprache, länger als ein einzelner Satz

**Literat** – Konzept für **schriftliches** Medium

- » literat-schriftliche Texte
- » Readability

**Oral** – Konzept für **akustisches** Medium

- » oral-akustische Texte
- » Listenability

Medien	Konzepte	
	literat	oral
schriftlich	Stummes Lesen eines Zeitungsartikels	Stummes Lesen einer Chatnachricht
akustisch	Anhören eines vorgelesenen Zeitungsartikels	Persönliches Gespräch

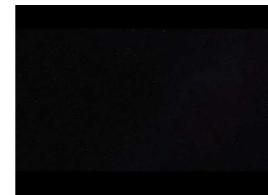
5 | Johanna Sacher, 4.2.2021

- Doch was ist überhaupt Diskurs?
- Diskurs wird von der Linguistik definiert als eine „Einheit von Sprache, die länger ist als ein einzelner Satz“
- ich unterscheide in meiner Arbeit zwischen literat-schriftlichem und oral-akustischem Diskurs
- literat und oral sind Konzepte, schriftlich und akustisch Medien
- und wenn ein Text für das schriftliche Medium konzipiert und anschließend auch über dieses Medium kommuniziert wird, dann kann er als „literat-schriftlich“ eingeordnet werden
- wie gut verständlich ein Text ist, der schriftlich übermittelt wird, wird „Readability“ genannt
- gleiches gilt für das orale Konzept, dieses hat als Zielmedium die akustische Kommunikation
- und wie gut ein über das akustische Medium kommunizierter Text ist, beschreiben wir mit der „Listenability“ dieses Textes
- natürlich können Medien und Konzepte auch anders kombiniert werden, bspw. wäre das stumme Lesen einer Chatnachricht ein orales Konzept, kommuniziert über das schriftliche Medium.
- genauso können literat konzipierte Texte, wie z.B. Zeitungsartikel, natürlich über das akustische Medium kommuniziert werden, bspw. wenn sie vorgelesen werden

# MOTIVATION

- Verwendung von Sprachassistenten zum Vorlesen von z.B. Zeitungsartikeln
- Zeitungsartikel wurde geschrieben, um gelesen zu werden  
» vorgelesen ggf. nicht mehr so gut verständlich

literates Konzept – akustisches Medium



literates Konzept – schriftliches Medium



By Dennis Overbye

Published Jan. 19, 2021 Updated Jan. 20, 2021



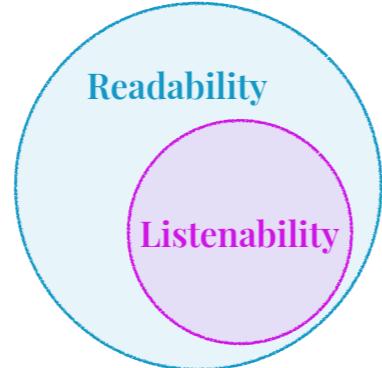
Astronomers are searching the cosmic lost-and-found for one of the biggest, baddest black holes thought to exist. So far they haven't found it.

In the past few decades, it has become part of astronomical lore, if not quite a law, that at the center of every luminous city of light, called a galaxy, lurks something like a hungry Beelzebub, a giant black hole into which the equivalent of millions or even billions of suns have disappeared. The bigger the galaxy, the more massive the black hole at its center.

6 | Johanna Sacher, 4.2.2021

- Was uns zum eigentlichen Ursprung dieser Arbeit führt
- Denn ganz am Anfang haben wir uns gefragt, was passiert, wenn Sprachassistenzen einen solchen Zeitungsartikel vorlesen, der eben für die Zeitung, also für das schriftliche Medium, konzipiert wurde
- Sprachassistenten können Dinge vorlesen, die wir zuvor selbst lesen mussten, z.B. Zeitungsartikel
- gibt uns als Nutzern freie Hände, um parallel ggf. andere Aufgaben zu erledigen oder einfach zu entspannen
- es kann aber sein, dass der Text auf diese Weise nicht mehr so gut verständlich ist
- das möchte ich kurz an einem Beispiel demonstrieren
- zuerst hören wir uns einen Satz aus einem Zeitungsartikel an, den Siri vorliest
- und dann können Sie den Satz selbst lesen und einmal vergleichen, was besser verständlich war
- Beispiele zeigen
- vielleicht war das Lesen jetzt auch einfacher, weil Sie den Satz zuvor schon gehört haben
- aber ich denke es ist deutlich, dass das Verstehen beim Zuhören schwieriger war als bei Lesen

LISTENABILITY & READABILITY



- **Readability** gut erforscht
  - » viele Methoden zur Messung
- **Listenability** kaum erforscht
  - » kaum Methoden zur Messung

**mündliche Tradition**



Quelle



**schriftliche Tradition**



Quelle

7 | Johanna Sacher, 4.2.2021

- das liegt daran, dass Listenability und Readability nicht das gleiche sind
- da Schrift und geschriebene Texte sich aus Sprache, aus Erzählungen entwickelt haben, ist ein Text mit hoher Listenability auch immer gut lesbar
- doch umgekehrt gilt das nicht
- und obwohl Readability gut erforscht ist und es viele Methoden gibt, sie zu messen,
- lässt sich zu Listenability kaum Forschung finden
- auch kaum Methoden zur Messung
- für die linguistischen Features werden häufig Readability Skalen verwendet
- das scheint für Listenability jedoch nicht auszureichen

Wie können Texte so formuliert werden,  
dass sie über beide Medien funktionieren?



## Welche Faktoren erhöhen die **Listenability** eines Textes?

8 | Johanna Sacher, 4.2.2021

- Wenn wir also erreichen wollen, dass ein Text über beide Medien funktioniert, dann müssen wir uns fragen,
- Welche Faktoren die Listenability eines Textes erhöhen
- dafür muss nicht unbedingt das ganze Konzept des Textes geändert werden, denn das orale Konzept beinhaltet ggf. Redundanzen und Füllwörter, die ein solcher Text nicht braucht.
- Aber wir wollen Eigenschaften finden, mit der die Listenability verbessert werden kann

## Forschung

- »Einfache« Texte [Flesch<sup>[4]</sup>, Harwood<sup>[4]</sup>, Chall & Dial<sup>[4]</sup>]
- Kurze Sätze [Ortmann & Dipper<sup>[4]</sup>]
- Wiederholungen [Ortmann & Dipper<sup>[4]</sup>]
- Koordination (u.a. Diskursmarker) [Ortmann & Dipper<sup>[4]</sup>]

## Ratgeber

- Einfache Wörter
- Kurze Sätze
- Wiederholungen
- Zahlen runden
- Bindewörter / Diskursmarker

- wirft man einen Blick in die existierende Forschung und die Ratgeberliteratur, findet man ein paar Empfehlungen:

- die Empfehlung zu „einfachen“ Texten, weil Listenability komplexer sei als Readability, finde sich in nahezu allen Forschungsarbeiten.
- viel konkreter wird es jedoch oft nicht
- 2019 haben Ortmann und Dipper in ihrer Arbeit jedoch festgestellt, dass orale Texte kürzere Sätze und mehr Wiederholungen enthalten, als schriftliche Texte
- außerdem enthalten sie mehr Koordination, bzw. Diskursmarker
- Journalisten und Schreiber fürs Radio oder Theater empfehlen zusätzlich noch einfach Wörter und das runden von komplizierten Zahlen
- Kurze Sätze und Wiederholungen lassen sich einfach realisieren, doch was ist mit Diskursmarkern?
- Es wird empfohlen, sie zu nutzen, um einen Flow im Text zu kreieren, wo im schriftlichen sonst z.B. Überschriften, Tabellen und Listen helfen
- Doch konkrete Regeln gibt es hier nicht
- die Satzlänge, die Schwierigkeit eines Wortes, die Anzahl an Wiederholungen lassen sich einfach messen, aber Diskursmarker sind nicht ganz so einfach zu verwenden
- und vor allem: Was sind eigentlich Diskursmarker?

# DISKURSMARKER

## Begriff

Nach Das et al., 2018 [1]  
(EnDimLex)

- lexikaler Ausdruck, kann nicht flektiert werden
- signalisiert zweiseitige Relation zwischen Diskurssegmenten
- feststehender, nicht modifizierbarer Ausdruck
  - ↳ nicht: *for this reason (for this exact reason)*
- nicht semantisch kombinierbar
  - ↳ nicht: *particularly if*
- feststehende Phrasen ok: *even if*

I love the Shire. I begin to wish, somehow, that I had gone, too.

I love the Shire. **But** I begin to wish, somehow, that I had gone, too.

10 | Johanna Sacher, 4.2.2021

- Begriff „Diskursmarker“ ist nicht eindeutig
- zum einen Verschiedene Begriffe:
  - Cue Phrase
  - Discourse Connective
  - Connective
- zum anderen ist der Begriff selbst nicht klar definiert:
  - gibt verschiedene Ansätze
  - Die in der Computerlinguistik verwendete Definition ist aber gut zusammengefasst mit den Kriterien, die 2018 von den Autoren der EnDimLex Liste aufgestellt wurden
  - Sie haben eine Liste Englischer Discourse Connectives erstellt, die frei zugänglich ist
  - um in die Liste aufgenommen zu werden, musste ein Wort folgende Kriterien erfüllen:
    - lexikaler Ausdruck, kann nicht flektiert werden
    - eine Beziehung zwischen zwei Diskurssegmenten signalisieren
    - es muss ein feststehender, nicht modifizierbarer Ausdruck sein (z.B. nicht „for this reason“, weil man das einfach zu „for this exact reason“ umformulieren kann)
    - außerdem darf es nicht semantisch kombinierbar sein, d.h. Kombinationen wie „particularly if“ sind nicht erlaubt.
    - Feststehende Phrasen wie „even if“ hingegen schon
- Da Diskursmarker im Prinzip eine funktionale Wortgruppe sind, schauen wir uns das an einem Beispiel an
- Dieser Satz ergibt zwar Sinn, wie er da steht, ist aber, vor allem ohne Kontext, nicht ganz Eindeutig:
  - Es könnte sein, dass Frodo woanders ist und wünscht, er wäre ins Auenland gegangen, oder andersherum
- Der Diskursmarker macht klar, wie die beiden Sätze miteinander in Beziehung stehen:
  - Frodo wünschte, er wäre mit Bilbo zu den Elben gegangen, obwohl er das Auenland liebt
- Der Diskursmarker dient uns also als Wegweiser im Text

Diskursmarker setzen sich aus verschiedenen anderen Wortgruppen zusammen  
» erschwert automatische Erkennung beim String-Matching

Gollum was very angry **as a result** of Bilbo winning the ring. (*Adverb*)

Bilbo won the ring. **As a result**, Gollum was very angry. (*Diskursmarker*)

- Die Diskursmarker wurden mit Hilfe von Stringmatching in den Texten identifiziert
- und ein Problem bei der korrekten Identifizierung von Diskursmarkern ist, dass sie eine funktionale Gruppe sind, die sich aus anderen Wortgruppen zusammensetzt
  - > u.a. aus Adverbien, Konjunktionen, Präpositionalphrasen
  - > d.h. für eine Erkennung könnte man z.B. mit Part-of-Speech Tagging nach dem Ausschlussverfahren vorgehen.
  - > Beim 1. Satz würde ein Adverb erkannt werden
  - > Beim 2. nicht. Dann müsste man die nicht identifizierten Worte und Phrasen noch mal genauer betrachten
  - > Für unsere Zwecke reicht das Ergebnis aber aus, auch wenn Homographen eines Diskursmarkers mitgezählt werden

# **DISKURSMARKER**

## **Zusammenfassung**

- keine inhaltliche Bedeutung
- signalisieren Beziehungen zwischen Diskurssegmenten
- funktionale Gruppe
  - » setzt sich aus verschiedenen Wortgruppen zusammen

12 | Johanna Sacher, 4.2.2021

-zusammenfassend:

- DM tragen keine inhaltliche Bedeutung zum Satz bei
- sie signalisieren vielmehr die Beziehung zwischen zwei Diskurssegmenten
- sie sind eine funktionale Wortgruppe, die sich aus anderen Wortgruppen wie Adjektiven, Konjunktionen und Präpositionalphrasen zusammensetzt

# DISKURSMARKER

## Bedeutungsgruppen

Nach Das et al., 2018 [1]  
(EnDimLex)

### COMPARISON Vergleich

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

### CONTINGENCY Folgern

*so, for, because, given, in case, whatever, ...*

### EXPANSION Hinzufügen eines Aspektes

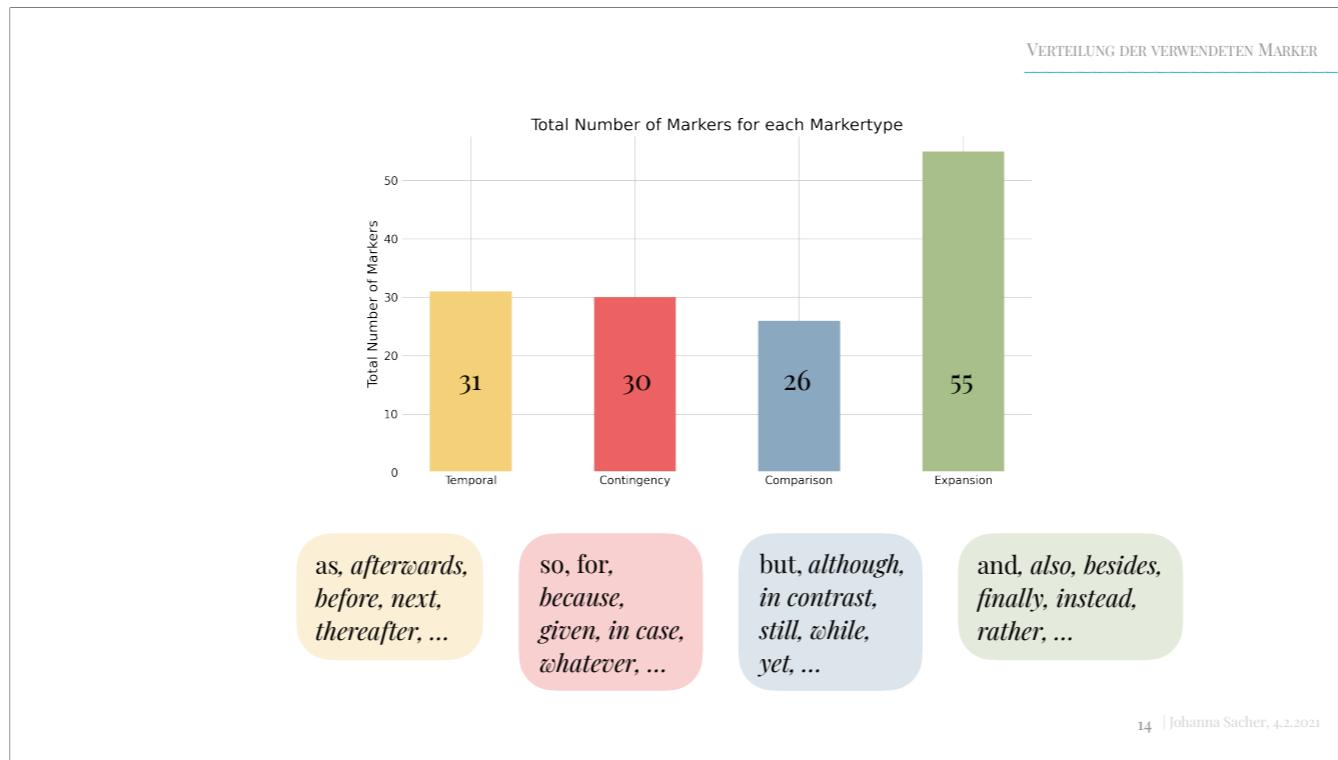
*and, also, besides, finally, instead, rather, ...*

### TEMPORAL Zeitlicher Bezug

*afterwards, as, before, next, thereafter, ...*

13 | Johanna Sacher, 4.2.2021

- Wir haben grade gesehen, dass DM eine funktionale Gruppe sind
- Bei dieser Funktion wollen wir ein bisschen ins Detail gehen
- und uns ansehen, welche Beziehung genau jeweils von den einzelnen Markern signalisiert wird
- Genau wie bei den verschiedenen Ansätzen zur Definition, gibt es auch verschiedene Ansätze für die Einteilung in Bedeutungsgruppen
- Die meisten lassen sich jedoch stimmen im Grunde miteinander überein, sie nennen die Gruppen nur unterschiedlich oder treffen detailliertere Unterscheidungen
- In unsere Betrachtung gibt es vier Gruppen
  - Comparison Marker, die einen Vergleich zwischen den Segmenten anzeigen (but, although, in contrast)
  - Contingency Marker, die anzeigen, dass ein Segment sich aus dem anderen folgern lässt (so, for, given)
  - Expansion Marker, die signalisieren, dass ein Aspekt hinzugefügt wird (and, also, besides)
  - und Temporal Marker, die die Segmente zeitlich miteinander in Bezug setzen (afterwards, as, before, next)
  -



- Betrachtet man die von uns verwendete Liste, ergibt sich für die enthaltenen Marker diese Verteilung auf die vier Klassen
- mit 55 der 149 Marker haben die Expansion Marker den weitaus größten Anteil
- zu dieser Gruppe gehören Wörter wie and, also, besides und finally
- die zweitgrößte Gruppe sind mit 31 Markern die Temporal Marker,
- dicht gefolgt von den 30 Contingency Markern
- und mit 26 Markern enthält die Comparison Klasse die wenigsten Marker

Diskursmarker können in mehrer Klassen gleichzeitig fallen:

Sam and Pippin crouched behind a large tree-bole, **while** Frodo crept back a few yards towards the lane.

(*Temporal & Comparison*)

**Since** they were all hobbits, and were trying to be silent, they made no noise that even hobbits would hear.

(*Contingency*)

I came also upon two others, but they turned away southward. **Since** then I have searched for your trail.

(*Temporal*)

- Ein Problem bei einer solchen Einteilung ist, dass ein DM je nach Kontext in verschiedene Klassen fallen kann, oder sogar in mehrer gleichzeitig
- Im ersten Satz: Zeitliche und Vergleich
- im 2. wird Since als Begründung verwendet
- im 3. jedoch im zeitlichen Sinne. Hier ist außerdem nicht 100% klar, ob Since (then) überhaupt ein DM ist oder nicht.
- Auch dieses Problem war für unsere Zwecke jedoch nicht so schlimm

# TEXTDATEN

## Corpora

akustische Corpora mit Transkripten von Audiomaterial &  
schriftliche Corpora mit ursprünglich schriftlichem Material

### KRITERIEN

kostenlos

groß

qualitativ hochwertig

nachrichtenähnlich

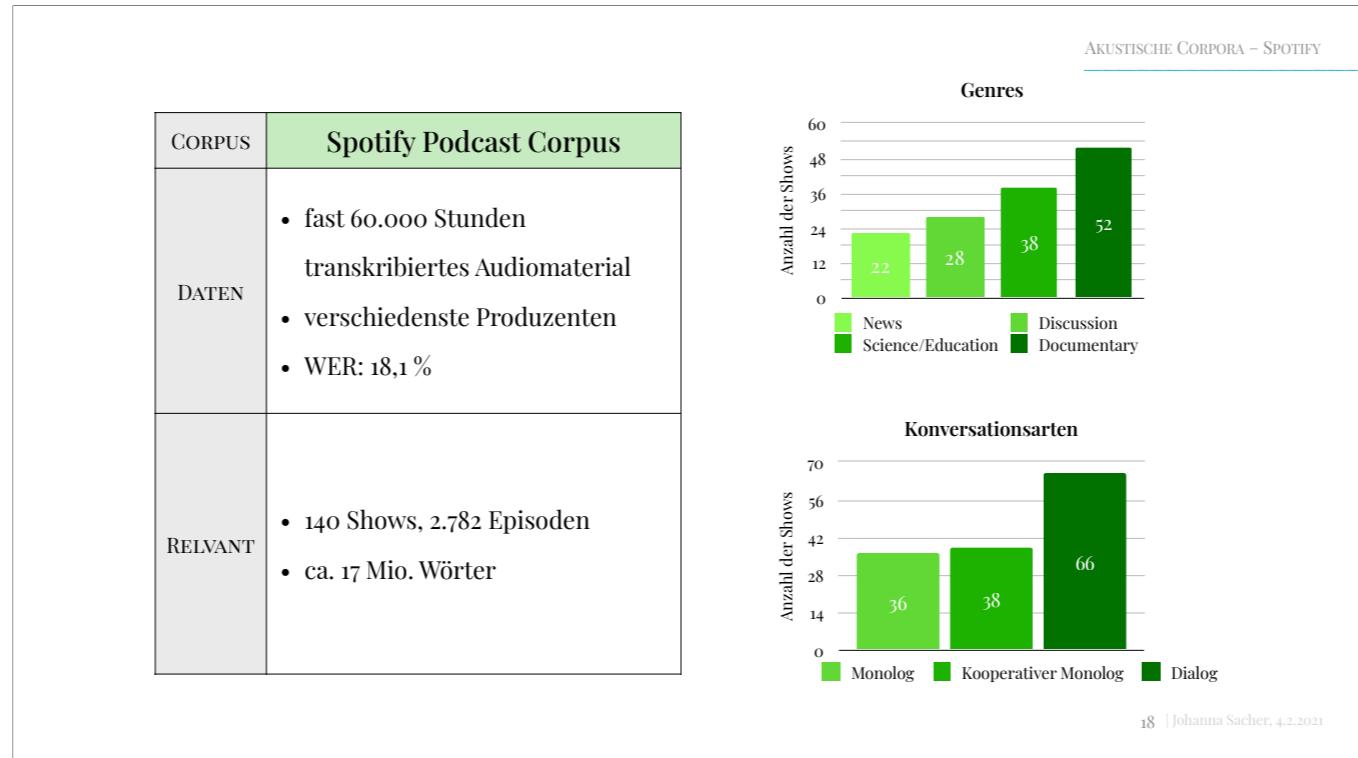
16 | Johanna Sacher, 4.2.2021

- Um das Vorkommen von Diskurmarkern auch tatsächlich untersuchen zu können, brauchten wir natürlich auch Texte
- da wir vergleichen wollten zwischen schriftlichem und akustischem Diskurs, brauchten wir also auch schriftliche und akustische Texte
- wobei akustische Texte Transkripte von Audiomaterial sind, und schriftliche eben Texte, die ursprünglich schriftlich sind.
- schriftliche Corpora waren bereits vorhanden, aber akustische Corpora mussten wir zuerst ausfindlich machen
- dabei war wichtig, dass sie kostenlos und groß waren,
- von möglichst hoher Qualität
- und nachrichtenähnlich, da wir uns am Anfang der Arbeit vor allem Conversational News konzentriert haben

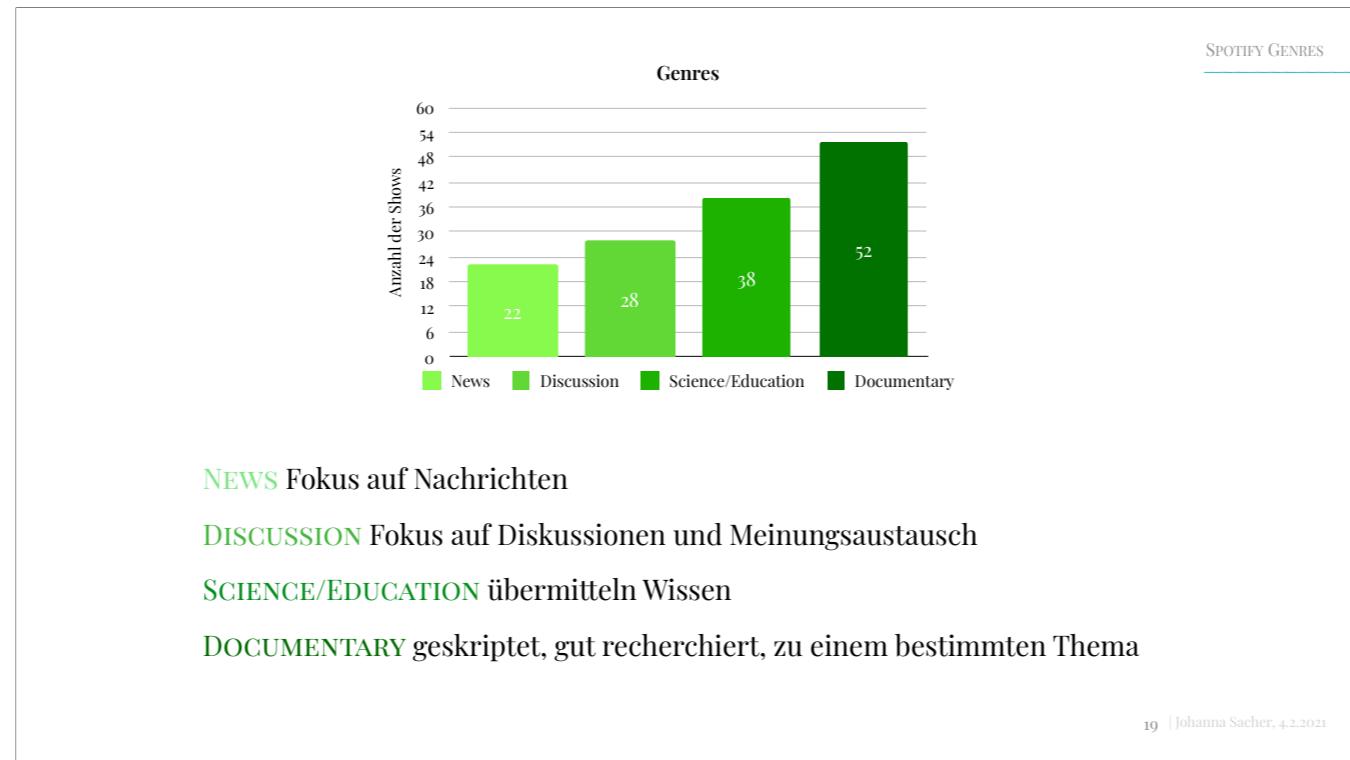
CORPUS	Spotify Podcast Corpus	TED-LIUM 3 Corpus
DATEN	<ul style="list-style-type: none"> <li>• fast 60.000 Stunden transkribiertes Audiomaterial</li> <li>• verschiedenen Produzenten</li> <li>• WER: 18,1 %</li> </ul>	<ul style="list-style-type: none"> <li>• 1.983 TED-Talks</li> <li>• ca. 4 Mio. Wörter</li> <li>• WER: 6,7 %</li> </ul>

17 | Johanna Sacher, 4.2.2021

- letztendlich haben wir zwei akustische Corpora gefunden, den Spotify Podcast Corpus und den TED-LIUM 3 Corpus
- der Spotify Corpus enthält 60.000 h transkribiertes Audiomaterial von diversen Spotify Podcasts und hat eine WER von 18.1%
- der TED Corpus besteht aus knapp 2000 TED-Talks mit ca. 4 Mio Wörtern.
- Er hat eine WER von 6,7 %



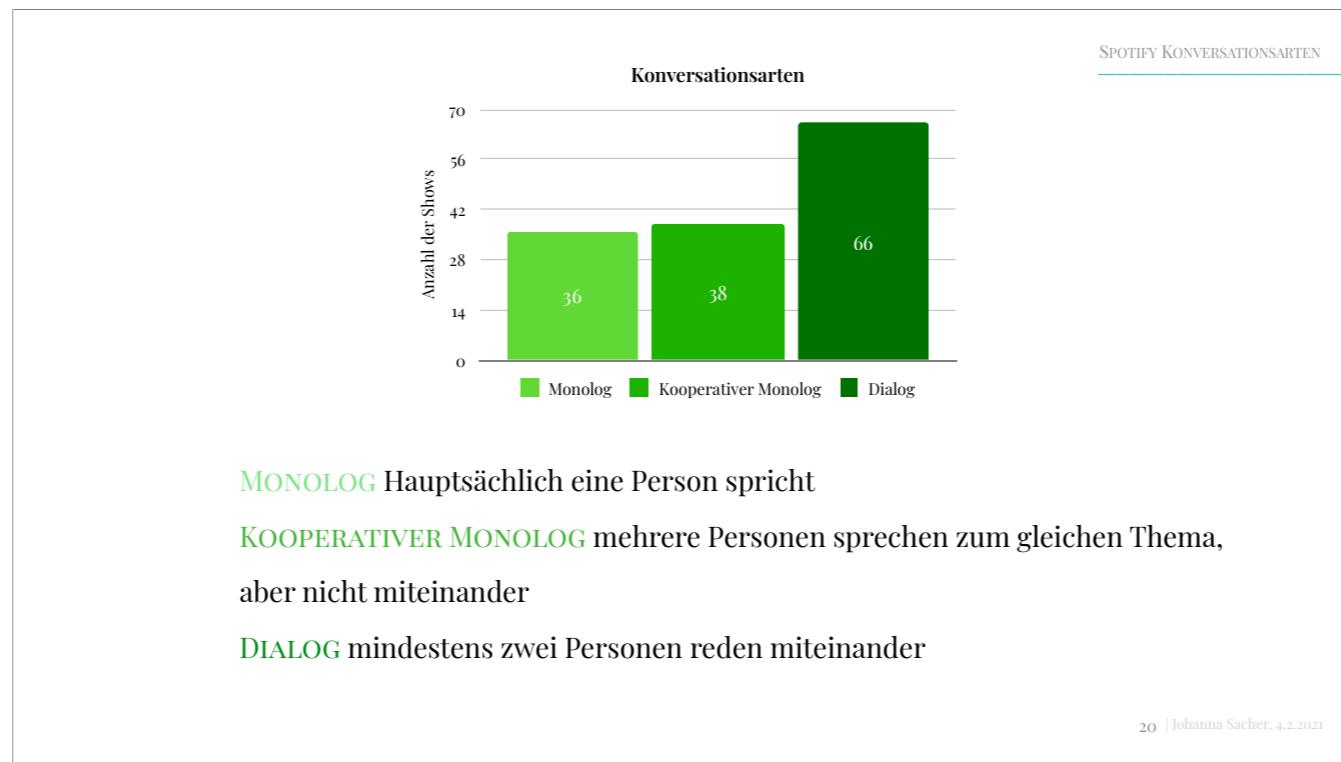
- Da der Spotify Corpus aus den unterschiedlichsten Podcasts besteht, habe ich das vorhandene Material zuerst nach Genres durchsucht
- Dann habe ich sachliche Genres wie verschiedene News, Documentary, History, Science und True Crime herausgesucht
- damit sind wir am Ende auf 140 Shows mit knapp 2.800 Episoden und ca. 17 Mio. Wörtern gekommen.
- Diese Shows habe ich dann in 4 Genres und 3 Konversationsarten unterteilt



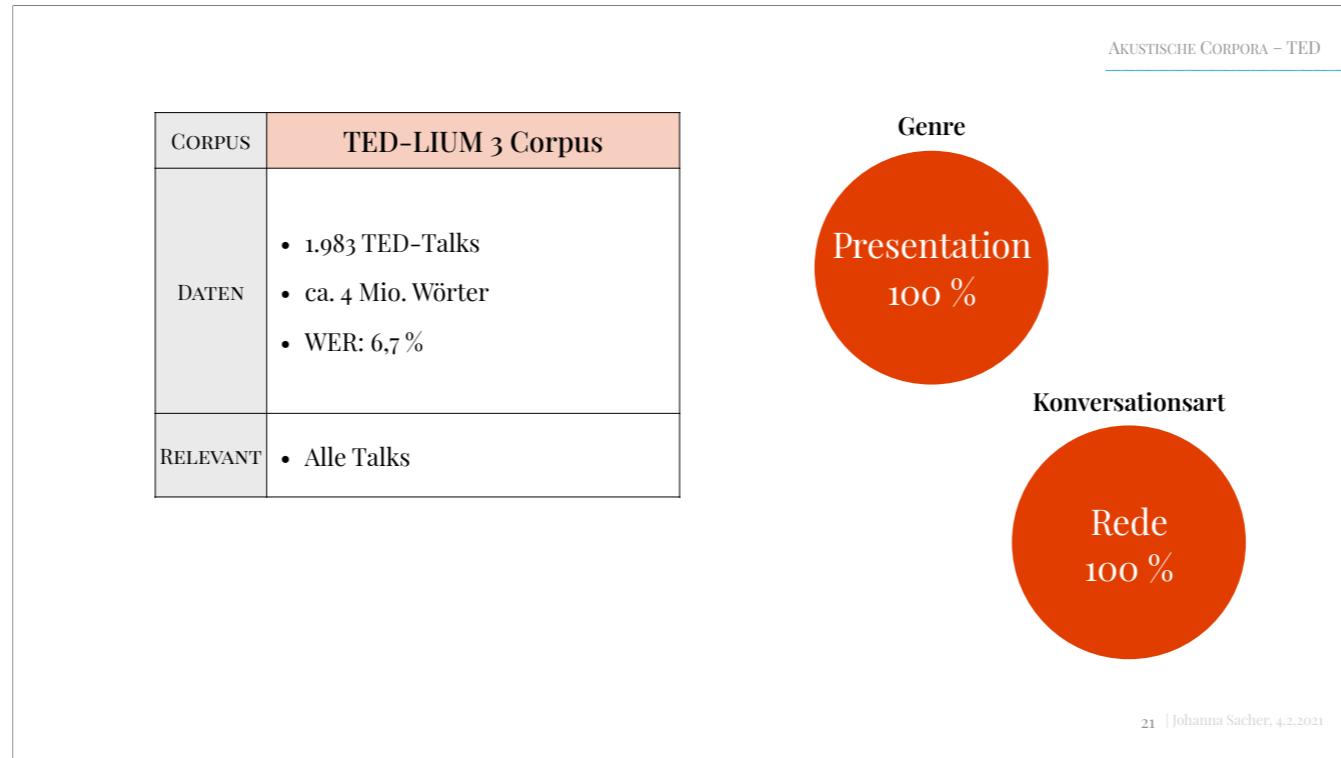
Die Genres sind

News, Discussion, Science/Education und Documentary

- und ich habe die Podcasts v.a. anhand ihrer Beschreibungen eingeteilt
- News: Fokus auf Nachrichten, egal zu welchem Thema (22 Shows)
- Discussion: Fokus liegt auf Diskussionen und Meinungsaustausch (28)
- Science/Education: Wissenschaftsshows oder Shows über sehr spezielle Themen, z.B. Podcasts von Fachleuten für andere Fachleute (38 Shows)
- Documentary: geskriptete, gut recherchierte Shows zu einem bestimmten Thema. Oft z.B. True Crime (52 Shows)



- neben Genres haben wir außerdem unterschieden nach der Art, wie der Diskurs stattfindet, also nach Konversationsarten
- hier konnte ich die Podcasts einteilen in Monolog, kooperativen Monolog und Dialog
- auch hierbei habe ich mich an den Beschreibungen orientiert, habe aber auch in fast alle Shows einmal reingehört, um ein besseres Bild zu bekommen
- Monolog bedeutet, dass hauptsächlich eine Person spricht (36 Shows)
- kooperativer Monolog, dass mehrere Personen sprechen zum gleichen Thema, aber nicht miteinander
  - > kommt z.B. oft in Dokumentationen vor, wenn Journalisten oder Sprecher sich zwischendurch abwechseln und über einen Teil des Themas reden, den sie recherchiert haben (38 Shows)
- im Dialog reden zwei oder mehr Personen miteinander (66 Shows)



- Der TED Corpus enthält keinerlei Informationen über Thema des Talks, nur den Redner und den Titel des jeweiligen Talks
- wir haben uns deshalb dazu entschieden, den Talks ein eigenes Genre und eine eigene Konversationsart zu geben, und denen dann jeweils alle Talks zuzuordnen
- TED Talks sind generell sehr speziell und haben ihren eigenen „Stil“



- wir haben also zusätzlich das Genre „Presentation“, bei dem etwas nach einem vorbereiteten Skript vor einer Menge an Zuhörern live präsentiert wird

### Konversationsart



TED – KONVERSATIONSART

REDE vor einer Menge Zuhörer nach einem  
vorbereiteten Skript gehalten

23 | Johanna Sacher, 4.2.2021

- und die Konversationsart Rede, die nach einem Skript live vor einer Menge Zuhörer gehalten wird

CORPUS	New York Times Corpus	Gigaword Corpus
DATEN	<ul style="list-style-type: none"> <li>• 1,8 Mio. Nachrichtenartikel der New York Times</li> <li>• ca. 1,1 Mrd. Wörter</li> </ul>	<ul style="list-style-type: none"> <li>• Newswire Textdaten</li> <li>• aus 7 Quellen</li> <li>• ca. 4 Mrd. Wörter</li> </ul>
RELEVANT	<ul style="list-style-type: none"> <li>• Alles</li> </ul>	<ul style="list-style-type: none"> <li>• Alles</li> </ul>

24 | Johanna Sacher, 4.2.2021

- Bei den schriftlichen Corpora war das ganze etwas einfacher, hier konnten wir an Nachrichtentexten einfach den New York Times Corpus verwenden,
- dieser hat 1,8 Mio. Artikel mit ca. 1.1 Mrd. Wörtern
- außerdem haben wir den Gigaword Corpus benutzt
- der ca. 4 Mrd. Wörter enthält
- diese beiden Corpora haben wir nicht nach Genres und Konversationsarten aufgeteilt, weil sie nur schriftliche Nachrichtenartikel enthalten

TYP	Akustische Corpora		Schriftliche Corpora		
	CORPUS	Spotify	TED-LIUM 3 Corpus	New York Times	Gigaword
RELEVANT		<ul style="list-style-type: none"><li>• 2.782 Episoden</li><li>• ca. 17 Mio. Wörter</li></ul>	<ul style="list-style-type: none"><li>• 1.983 TED-Talks</li><li>• ca. 4 Mio. Wörter</li></ul>	<ul style="list-style-type: none"><li>• 1,8 Mio. Artikel</li><li>• 1,1 Mrd. Wörter</li></ul>	<ul style="list-style-type: none"><li>• ca. 4 Mrd. Wörter</li></ul>

# TEXTSORGEN

## Diskursarten

- oral-akustisch
- literat-schriftlich

## Genres *(oral-akustisch)*

- News
- Discussion
- Science/Education
- Documentary
- Presentation

## Konversationsarten *(oral-akustisch)*

- Dialog
- Monolog
- Kooperativer Monolog
- Rede

26 | Johanna Sacher, 4.2.2021

- Wir können jetzt also verschiedene Textsorten miteinander vergleichen
- zum einen haben wir oral-akustische und literat-schriftliche Texte.
  - die oral-akustischen sind die Texte des Spotify und des TED Corpus
  - die literat-schriftlichen die des New York Times und des Gigaword Corpus
- Dann wurden die oral-akustischen Texte auf zwei verschiedene Arten aufgeteilt
  - einmal nach Genres (News, Discussion, Science/Education, Documentary, Presentation)
  - und einmal nach Konversationsarten (Dialog, Monolog, Kooperativer Monolog, Rede)
- Bei diesen Textsorten gibt es außerdem interaktiver, wie z.B. den Dialog, und passiver, wie den Monolog oder den kooperativen Monolog
- es gibt sachliche Genres, wie z.B. News und Science oder Documentary, und es gibt interaktionslastigere, wie Discussion und Presentation oder Rede, bei denen mit Diskurspartnern oder den Zuhörern interagiert wird

# FRAGEN

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?

27 | Johanna Sacher, 4.2.2021

- Womit wir zur eigentliche Analyse gehen können
- Konzentriert habe ich mich dabei auf fünf Fragen:

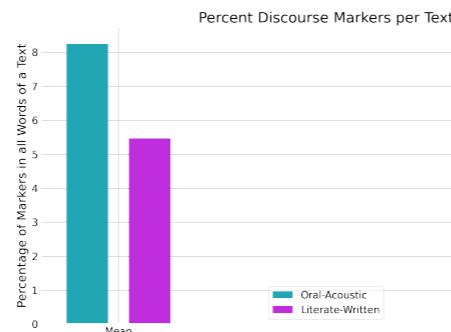
Welche Textsorten stützen sich besonders auf Diskursmarker?  
An welchen Positionen im Text stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?  
An welchen Positionen im Satz stützen sich die jeweiligen Textsorten besonders auf Diskursmarker?  
Auf welche Klassen von Diskursmarkern stützen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

Details und Plots zu jeder Frage und jeder Textsorten-Kategorie gibt es in der Arbeit  
aber ich möchte einen Überblick über die wichtigsten oder auffälligsten Ergebnisse geben

# AUSWERTUNG

## 1. Generelle Verteilung

- improvisierte oral-akustische Texte nutzen mehr Diskursmarker als geskriptete
- oral-akustische Texte nutzen mehr Diskursmarker als literat-schriftliche



P-Wert < 0.001

» **oral-akustische** Texte nutzen mehr  
Diskursmarker als **literat-schriftliche**

28 | Johanna Sacher, 4.2.2021

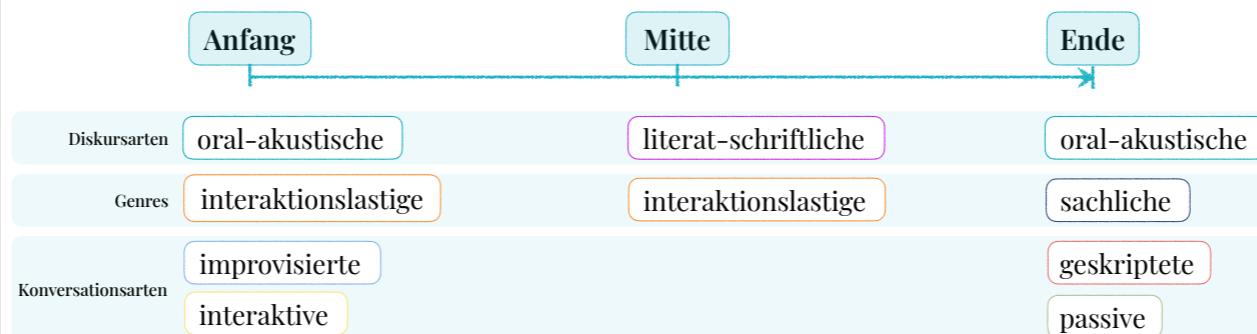
Die erste Frage befasst sich mit der generellen Verteilung

- hier wurde klar deutlich, dass OA Texte mehr DM nutzen als LS Texte
- Im Diagramm sieht man den durchschnittlichen prozentualen Anteil von Diskursmarkern an allen Wörtern der jeweiligen Texte,
- türkis abgebildet sind die OA Texte, pink die LS Texte
- man kann deutlich sehen, dass die OA Texte mit über 8% durchschnittliche mehr Diskursmarker nutzen, als LS Texte.
- diese haben durchschnittlich nur einen Anteil von 5% Diskursmarkern
- dieser Unterschied ist, wie alle hier vorgestellten Ergebnisse, statistisch signifikant
- interessant in diesem Zusammenhang ist außerdem, dass improvisierte OA Texte, z.B. Diskussionen, mehr DM nutzen als geskriptete, wie z.B. Dokumentationen

# AUSWERTUNG

## 2. Textpositionen

Texte, die an der jeweiligen Position im Text je **mehr Diskursmarker** nutzen als andere



29 | Johanna Sacher, 4.2.2021

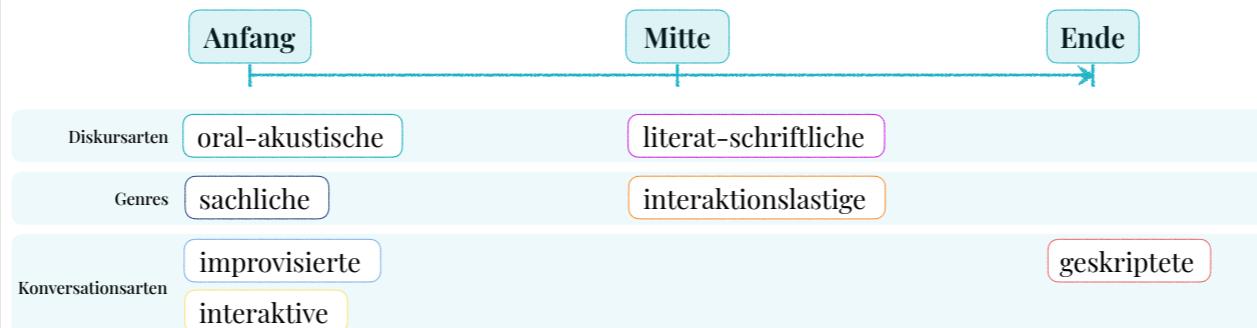
Die zweite Frage bezieht sich auf die Position von Diskursmarkern im Text

- zum einen haben wir festgestellt, dass Diskursmarker in allen Texten sehr gleichmäßig über den gesamten Text verteilt sind
- im Vergleich wurde deutlich, dass oral-akustische Texte am Textanfang und Ende mehr DM nutzen als LS Texte.
- LS Texte nutzen dafür in der Mitte mehr DM als OA
- außerdem nutzen interaktionslastige, interaktive und improvisierte Texte am Anfang mehr DM als andere,
- während sachliche, geskriptete und passive Texte sie eher am Ende nutzen

# AUSWERTUNG

## 3. Satzpositionen

Texte, die an der jeweiligen Position im Satz je **mehr Diskursmarker** nutzen als andere



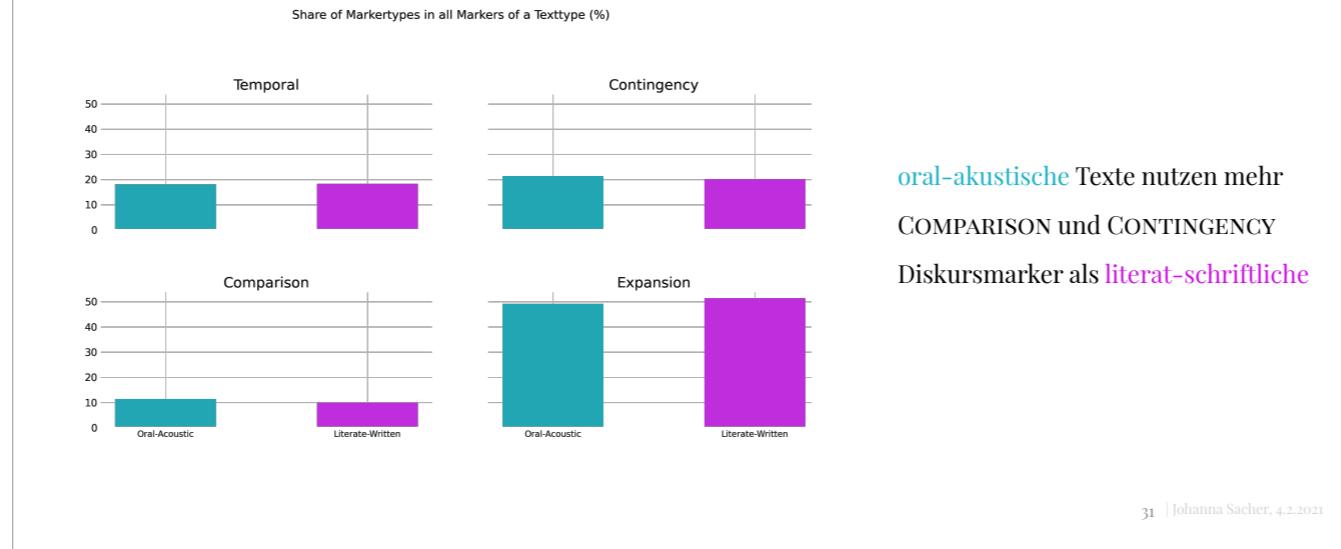
30 | Johanna Sacher, 4.2.2021

Bei der dritten Frage ging es um die Position der Diskursmarker im Satz

- innerhalb der Sätze wurden in allen Texten Diskursmarker ein bisschen häufiger am Anfang des Satzes verwendet als am Ende.
- im Vergleich nutzen OA Texte am Anfang des Satzes mehr DM als LS Texte
- LS Texte nutzen dafür in der Mitte des Satzes mehr DM als OA Texte
- sachlichere Genres beginnen ihre Sätze eher mit DM als interaktionslastigere
- und improvisierte und interaktive Konversationsarten wie Dialoge beginnen ihre Sätze eher mit Diskursmarkern als geskriptete, passive, wie Texte des kooperativen Monologs

# AUSWERTUNG

## 4. Diskursmarker-Klassen



- Dann haben wir noch die Nutzung der vier Diskursmarker Klassen in den einzelnen Textsorten verglichen
- Temporal und Expansion Marker werden von oral-akustischen und literat-schriftlichen Texten etwa gleich häufig verwendet
- während oral akustische Texte wesentlich mehr Comparison und Contingency Marker einsetzen als literat-schriftliche
- das könnte damit zusammenhängen, dass einfache, schriftliche Nachrichten überlegt formulieren können, was gesagt werden soll,
- während improvisierte Texte oder Texte, die einem Spannungsbogen folgen, wie Reden oder Dokumentationen viele Vergleiche anstellen.
- und z.B. Science Shows und Diskussionen ziehen häufig Schlussfolgerungen

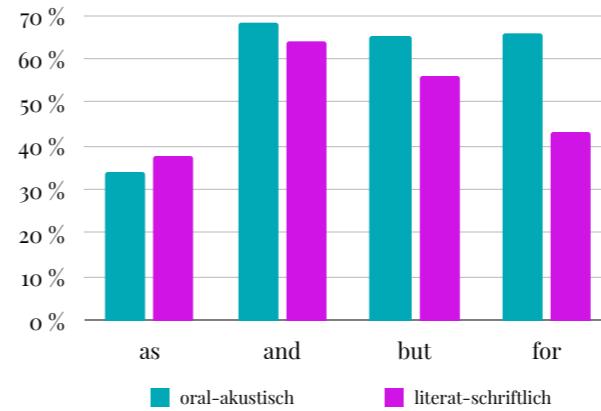
## AUSWERTUNG

### 5. Häufigste Diskursmarker der Klassen

as TEMPORAL  
and EXPANSION  
but COMPARISON  
for CONTINGENCY

Diskursarten

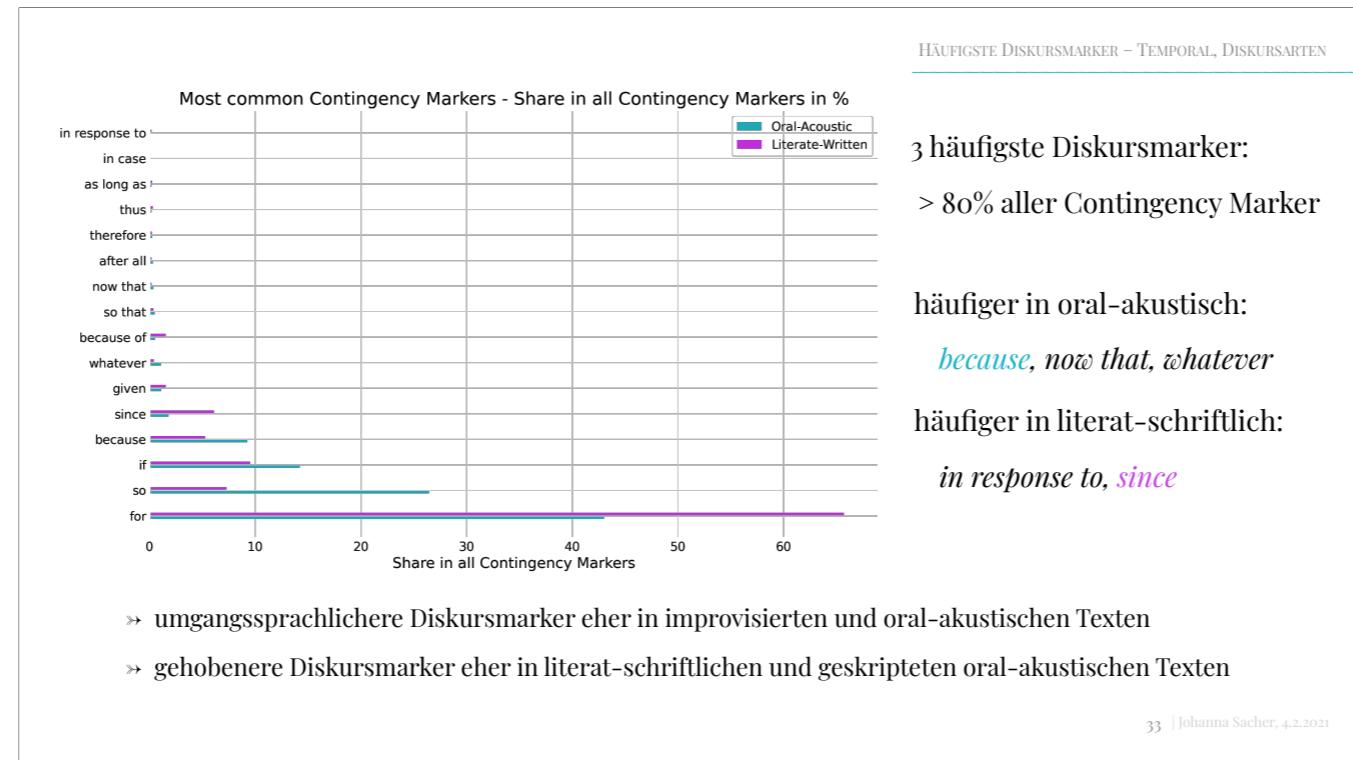
prozentualer Anteil der Marker an den Markern ihrer jeweiligen Klasse



32 | Johanna Sacher, 4.2.2021

in der 5. Frage haben wir die einzelnen Marker betrachtet, bzw. welche Marker in den jeweiligen Klassen die häufigsten sind

- Dabei konnten wir feststellen, dass die drei häufigsten Marker in jeder Klasse für jede Textsorte mehr als 70% aller benutzten Marker dieser Klasse ausmachen
- Interessant ist auch, dass die häufigsten Marker für nahezu alle Textsorten dieselben sind
- und zwar ist der häufigste Temporal Marker as
- der häufigste Expansion Marker ist and
- bei den Comparison Markern ist es but
- und für Contingency for
- Dieses Diagramm zeigt, wie häufig diese Marker in den Diskursarten je vorkommen,
- wir sehen z.B. dass „and“ bei den oral-akustischen Texten einen Anteil von 68% an allen Expansion Markern hat,
- bei den literat-schriftlichen sind es 64%.
- auch but und for machen bei den oral-akustischen Texten jeweils mehr als die Hälfte aller Diskursmarker ihrer Klasse aus
- Es ist aber auch zu beachten, dass sie wahrscheinlich nicht an jeder Stelle als Diskursmarker benutzt wurden, das bleibt für kommende Arbeiten herauszufinden



Dieses Diagramm zeigt die häufigsten Contingency Marker für oral-akustische und literat-schriftliche Texte

- hier ist noch mal gut zu sehen, dass die 3 häufigsten Marker, for, so und if, insgesamt mehr als 80% der verwendeten Contingency Marker ausmachen
- was hier außerdem deutlich wird, ist der Unterschied in den verwendeten Markern
- So kommen because, now that und whatever wesentlich häufiger in OA Texten vor als in LS,
- in response to und since jedoch wesentlich häufiger in LS Texten
- insgesamt ergibt sich das Bild, dass umgangssprachlichere Diskursmarker wie eben because und whatever häufiger in OA Texte vorkommen
- und dort auch häufiger in improvisierten Texten
- etwas gehobenere Ausdrücke wie „since“ oder z.B. auch „although“ und „nonetheless“ kommen dagegen häufiger literat-schriftlichen Texten vor, oder auch in geskripteten oral-akustischen Texten

# ZUSAMMENFASSUNG

## AUSBlick

- Aussortieren von Homographen
- Akustischen Nachrichten-Corpus verwenden
- Listenability aller betrachteten Texte messen und mit Nutzung von Diskursmarkern in Zusammenhang setzen

## WICHTIG

Diskursmarker werden in verschiedenen Textsorten unterschiedlich eingesetzt

» Auswirkung auf Listenability

- oral-akustische Texte nutzen mehr Diskursmarker als literat-schriftliche
- improvisierte Texte nutzen mehr Diskursmarker als geskriptete
- interaktive Texte nutzen mehr Diskursmarker als passive

34 | Johanna Sacher, 4.2.2021

- Zu tun bleibt also das Aussortieren von Mehrdeutigkeiten
  - also z.B. mit Part-of-Speech Tagging herausfinden, ob ein Wort oder eine Phrase im Kontext ein Diskursmarker ist oder nicht
  - dann wäre es außerdem interessant, noch einmal einen akustischen Corpus zu verwenden, der richtige Nachrichtensendungen enthält, z.B. geskriptete Radionachrichten oder Podcasts von Verlagshäusern, die ihre schriftlichen Artikel auch als Podcasts veröffentlichen
  - und was auf jeden Fall gut wäre, wäre die Listenability aller betrachteten Texte mit einer geeigneten Methode zu messen und die Ergebnisse mit der Nutzung von Diskursmarkern in Zusammenhang zu stellen
  - wir sind davon ausgegangen, dass Listenability zum einen durch das orale Konzept von improvisierten, frei gesprochenen Podcasts gegeben ist, und dass zum anderen die Journalisten, die geskriptete Dokumentationen erstellen und die Redner der TED-Talks etwas von ihrem Handwerk verstehen
  - es wäre sehr spannend, das mit tatsächlichen Zahlen belegen zu können
- 
- Wir können also festhalten, dass Diskursmarker in verschiedenen Textsorten unterschiedlich eingesetzt werden, mit entsprechenden Auswirkungen auf die Listenability
  - so nutzen oral-akustische Texte durchschnittlich mehr Diskursmarker als literat-schriftliche
  - improvisierte Texte nutzen mehr als geskriptete
  - und interaktivere Texte nutzen mehr Diskursmarker als passive
  - diese unterschiedliche Verwendung sollte beachtet werden, wenn Tools für die Analyse und die Erstellung von Texten für verschiedene Medien benutzt werden

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

[1] Debopam Das, Tatjana Scheer, Peter Bourgonje, and Manfred Stede. *Constructing a lexicon of English discourse connectives*, 2018. <https://www.aclweb.org/anthology/W18-5042/>

[2] Kenneth A. Harwood, *Listenability and Readability*, 1955. <https://doi.org/10.1080/0363775509375133>

[3] Rudolf Flesch, *Marks of Readable Style: A Study in Adult Education*, 1943. <https://books.google.de/books?id=ZQX0zQEACAAJ>

[4] Jeanne S. Chall and Harold E. Dial, *Predicting listener understanding and interest in newscasts*, 1948. <https://www.jstor.org/stable/1473082?seq=1>

[5] Katrin Ortmann and Stefanie Dipper, *Variation between different discourse types: Literate vs. oral*, 2019. <https://www.aclweb.org/anthology/W19-1407/>