Diskursmarker

in

schriftlichem & akustischem Diskurs

Bachelorverteidigung

Johanna Sacher, 4.2.2021

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 Diskursmarkern in den genannten Textsorten nachgewiesen

Außerdem:

- Was sind Diskursmarker
- Problematiken
- Bedeutungsgruppen

Begriff

Diskurs

Einheit von Sprache, länger als ein einzelner Satz

Quelle

Diskursmarker

Wörter wie *and*, *but* und *so*

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

Literat vs. Oral

Literat – Konzept für das **schriftliche** Medium

- literat-schriftliche Texte (LS)
- Readability

Oral – Konzept für das **akustische** Medium

- oral-akustische Texte (OA)
- Listenability

Medien	Konzepte	
	literat	**oral**
schriftlich	Stummes lesen eines Zeitungsartikels	Stummes lesen eines YouTube Kommentars
akustisch	Anhören eines vorgelesenen Zeitungsartikels	Persönliches Gespräch

Wieso dieses Thema?

Verwendung von Sprachassistenten zum Vorlesen von z.B. Zeitungsartikeln

Zeitungsartikel wurde geschrieben, um gelesen zu werden

↠ vorgelesen ggf. nicht mehr so gut verständlich

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

Welche Faktoren erhöhen die Listenability eines Textes?

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

Begriff

- Begriff ist nicht eindeutig definiert
- Verschiedene Begriffe in Benutzung

Funktionale Definiton nach Bruce Fraser

- DM ist ein lexikaler Ausdruck
- In `<S1 S2>` muss ein DM Teil von `S2` sein
- DM trägt nicht zur semantischen Bedeutung der Sequenz bei, sondern signalisiert eine Relation zwischen `S1` und `S2`

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

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

You are the master of Bag End now. And also, I fancy, you'll find a golden ring.

You are the master of Bag End now. You'll find a golden ring.

Kriterien des EnDimLex

- DM ist ein lexikaler Ausdruck und kann nicht flektiert werden
- DM signalisiert eine zweiseitige Relation, deren Argumente abstrakte Objekte sind
- Argumente können in Klausel-, Satz- oder Phrasenstruktur ausgedrückt werden

Weitere Bedingungen

- feststehender, nicht modifizierbarer Ausdruck
  - ↠ nicht: *for this reason* (*for this exact reason*)

- darf nicht semantisch kombinierbar sein
  - ↠ nicht: *particularly if*
  - ↠ feststehende Phrasen sind ok: *even if*

Vergleich

Funktionale Definition (Fraser)		EnDimLex-Kriterien
lexikaler Ausdruck	✓	lexikaler Ausdruck
	→	kann nicht flektiert werden
signalisiert Relation zwischen Diskurssegmenten	✓	signalisiert zweiseitige Relation zwischen Klauseln, Sätzen oder Phrasen
trägt nicht zur Bedeutung des Satzes bei	←	
meistens Adverbien, Konjunktionen, Präpositionalphrasen	✓	meistens Adverbien, Konjunktionen, Präpositionalphrasen

DM setzen sich aus verschiedenen anderen Wortgruppen zusammen

↠ erschwert automatische Erkennung

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

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

Zusammenfassung

- keine inhaltliche Bedeutung
- signalisieren Beziehungen zwischen Diskurssegmenten
- setzen sich aus verschiedenen Wortgruppen zusammen
- automatische Erkennung schwierig

Bedeutungsgruppen

Gibt verschiedene Ansätze, DM anhand ihrer Bedeutung in Klassen aufzuteilen

Fraser

Contrastive Markers Kontrast zwischen S1 und S2

but, *alternatively, although, even so, still, yet, …*

Elaborative Markers Genauere Ausführung von S1 in S2

and, *also, besides, for instance, moreover, similarly, …*

Inferential Markers S2 kann aus S1 gefolgert werden

so, *consequently, therefore, thus, …*

EnDimLex

Comparison Vergleich

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

Contingency Folgern, Möglichkeiten aufzeigen

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, …

EnDimLex	Fraser	Funktion
Comparison	Contrastive	Vergleich, Kontrast
Contingency	Inferential	Folgern
Expansion	Elaborative	Ausführen, Illustrieren
Temporal	Elaborative, Inferential	Zeitlich in Bezug setzen

DM 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)

Diskursarten

- oral-akustisch
- literat-schriftlich

Genres

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

Konversationsarten

- Dialog
- Monolog
- Kooperativer Monolog
- Rede

Corpora

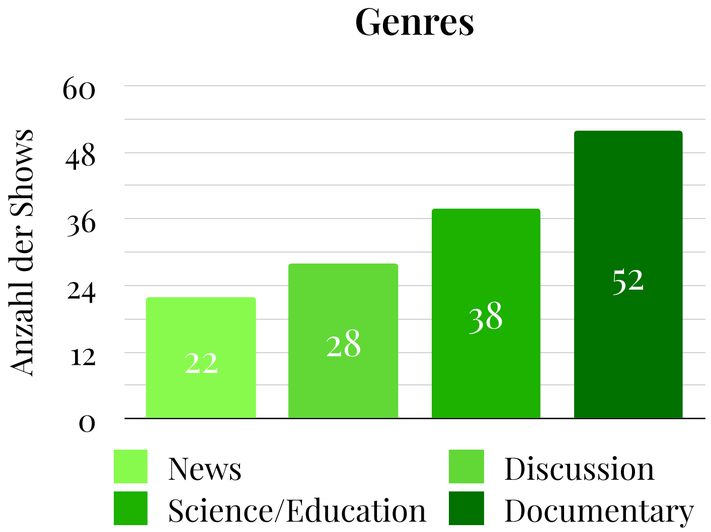
akustische Corpora mit Transkripten von Audiomaterial &

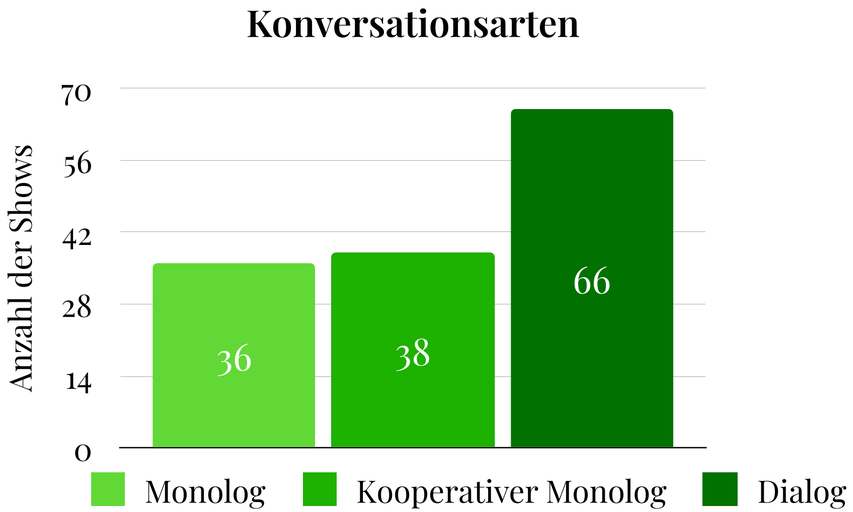
schriftliche Corpora mit ursprünglich schriftlichem Material

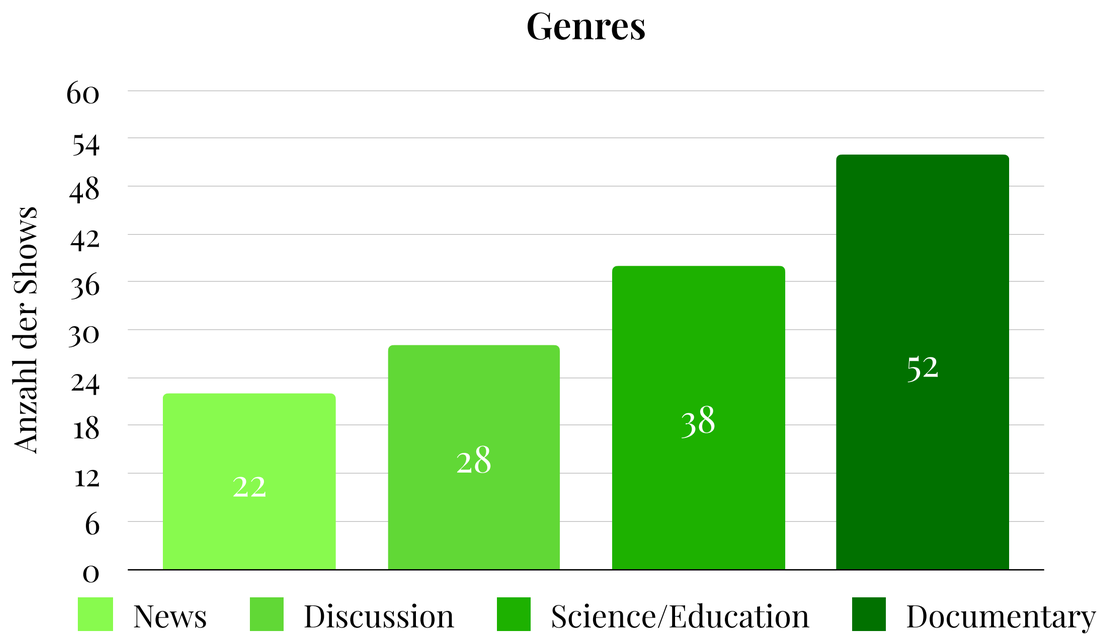
- kostenlos
- groß
- qualitativ hochwertig
- nachrichtenähnlich

Corpus	Spotify Podcast Corpus	TED-LIUM 3 Corpus
Daten	• fast 60.000 Stunden transkribiertes Audiomaterial • verschiedenste Produzenten • WER: 18,1 %	• 1.983 TED-Talks • ca. 4 Mio. Wörter • WER: 6,7 %

Corpus	Spotify Podcast Corpus
Daten	• fast 60.000 Stunden transkribiertes Audiomaterial • verschiedenste Produzenten • WER: 18,1 %
Nutzbar	• 140 Shows, 2.782 Episoden • ca. 17 Mio. Wörter





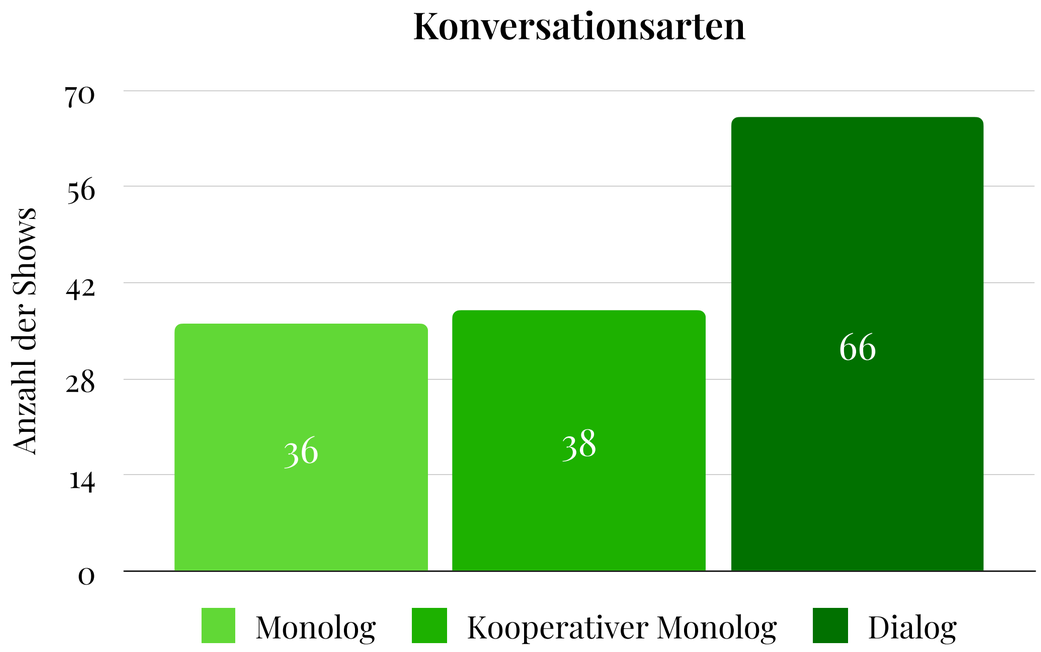


News Fokus auf Nachrichten

Discussion Fokus auf Diskussionen und Meinungsaustausch

Science/Education übermitteln Wissen

Documentary geskriptet, gut recherchiert, zu einem bestimmten Thema



Monolog Hauptsächlich eine Person spricht

Kooperativer Monolog mehrere Personen sprechen zum gleichen Thema, aber nicht miteinander

Dialog mindestens zwei Personen reden miteinander

Corpus	TED-LIUM 3 Corpus
Daten	• 1.983 TED-Talks • ca. 4 Mio. Wörter • WER: 6,7 %
Nutzbar	• Alle Talks

Genre

Konversationsart

Genre

Presentation vor einer Menge Zuhörer nach einem vorbereiteten Skript präsentiert

Konversationsart

Rede vor einer Menge Zuhörer nach einem vorbereiteten Skript gehalten

Corpus	Spotify Podcast Corpus	TED-LIUM 3 Corpus
Daten	• fast 60.000 Stunden transkribiertes Audiomaterial • verschiedenste Produzenten • WER: 18,1 %	• 1.983 TED-Talks • ca. 4 Mio. Wörter • WER: 6,7 %
Nutzbar	• 140 Shows, 2.782 Episoden • ca. 17 Mio. Wörter	• Alle Talks

Corpus	New York Times Corpus
Daten	• 1,8 Mio. Nachrichtenartikel der New York Times • ca. 1,1 Mrd. Wörter
Nutzbar	• Alles

Corpus	New York Times Corpus	Gigaword Corpus
Daten	• 1,8 Mio. Nachrichtenartikel der New York Times • ca. 1,1 Mrd. Wörter	• Newswire Textdaten • aus 7 Quellen • ca. 4 Mrd. Wörter
Nutzbar	• Alles	• Alles

Typ	Akustische Corpora		Schriftliche Corpora	
Corpus	Spotify	TED-LIUM 3 Corpus	New York Times	Gigaword
Nutzbar	• 140 Shows, 2.782 Episoden • ca. 17 Mio. Wörter	• 1.983 TED-Talks • ca. 4 Mio. Wörter	• 1,8 Mio. Artikel • 1,1 Mrd. Wörter	• ca. 4 Mrd. Wörter

Diskursmarker im Text erkennen

Diskursmarker wurden mit Hilfe eines einfachen

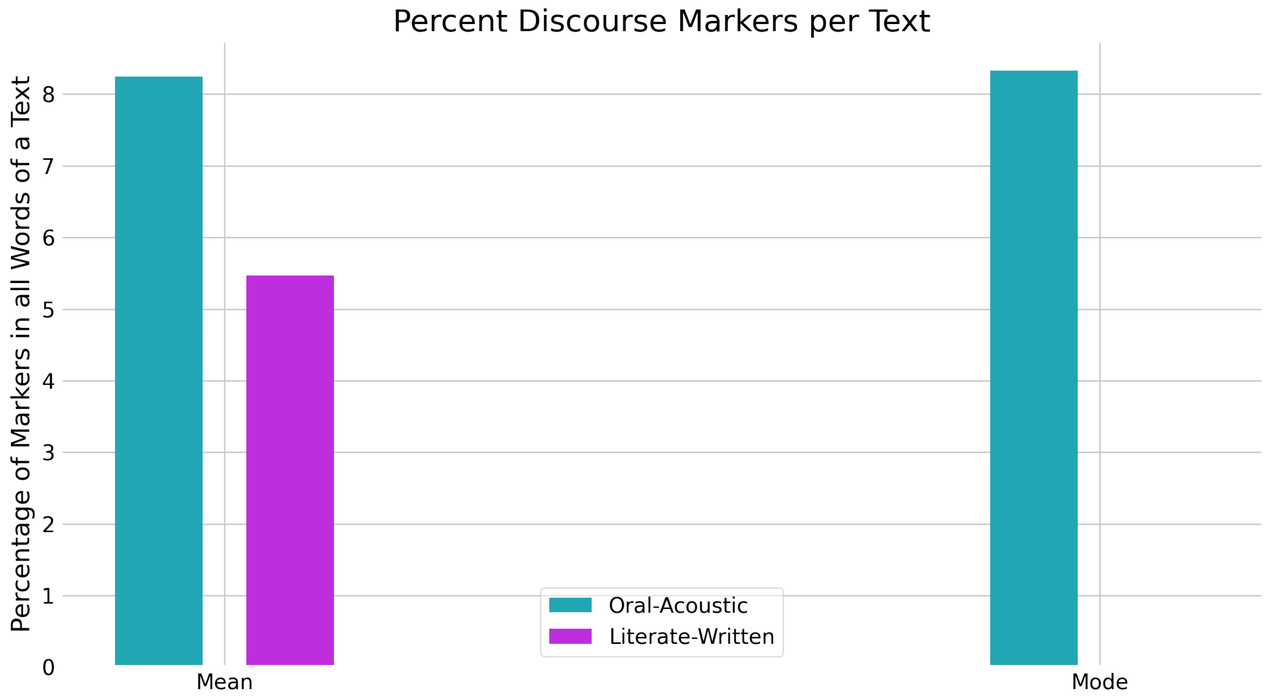
String-Matching Verfahren mit den Texten gematched

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?

1. Generelle Verteilung

- oral-akustische mehr als literat-schriftliche
- improvisierte OA Texte mehr als geskriptete



P-Wert < 0.001

↠ OA nutzt mit einer Effektgröße (EG) von 1,63 mehr DM als LS

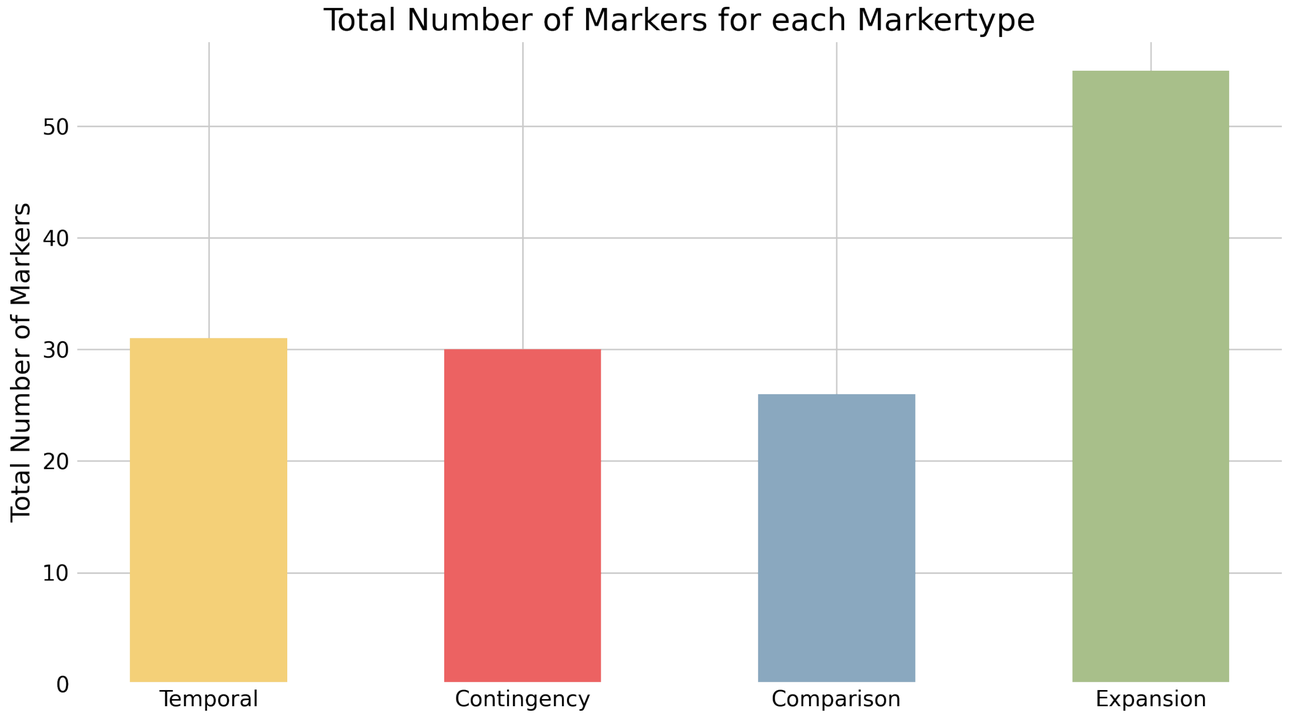
2. Textpositionen

- OA mehr am Anfang und am Ende, LS mehr in der Mitte
- interaktionslastigere Texte mehr am Anfang und in der Mitte, sachlichere mehr am Ende
- improvisierte mehr am Anfang, gespkriptete mehr am Ende

3. Satzpositionen

- OA mehr am Anfang, LS mehr in der Mitte
- sachliche Genres (News, Science) mehr am Anfang und weniger in der Mitte
- interaktionslastigere (Dialog) mehr am Anfang und weniger am Satzende, passive (Koop. Monolog) mehr am Satzende

4. Diskursmarker-Klassen



größter Anteil an allen DM in allen Texten: EXPANSION DM (u.a. *and*)

5. Häufigste Diskursmarker

Offene Fragen

Was bleibt zu tun?

Zusammenfassung

Überblick über die wichtigsten Ergebnisse