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

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

Diskursmarker

Wörter wie *and*, *but* und *so*

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

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

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.

mündliche Tradition

Quelle

schriftliche Tradition

Quelle

- Readability gut erforscht
  - ↠ viele Methoden zur Messung
- Listenability kaum erforscht
  - ↠ kaum Methoden zur Messung

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

Welche Faktoren erhöhen die Listenability eines Textes?

Forschung

- »Einfache« Texte [Flesch[3], Harwood[2], Chall & Dial[4]]
- Kurze Sätze [Ortmann & Dipper [5]]
- Wiederholungen [Ortmann & Dipper [5]]
- Koordination (u.a. Diskursmarker) [Ortmann & Dipper [5]]

Ratgeber

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

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)*

Zusammenfassung

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

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, …

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

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

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

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

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)

Corpora

akustische Corpora mit Transkripten von Audiomaterial &

schriftliche Corpora mit ursprünglich schriftlichem Material

KRITERIEN

- 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 %
Relvant	• 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 %
Relevant	• Alle Talks

Genre

Presentation
100 %

Konversationsart

Rede
100 %

Genre

Presentation vor einer Menge Zuhörer nach einem vorbereiteten Skript präsentiert

Konversationsart

Rede vor einer Menge Zuhörer nach einem vorbereiteten Skript gehalten

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
Relevant	• Alles	• Alles

Typ	Akustische Corpora		Schriftliche Corpora	
Corpus	Spotify	TED-LIUM 3 Corpus	New York Times	Gigaword
Relevant	• 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

Textsorten

Diskursarten

- oral-akustisch
- literat-schriftlich

Genres

(oral-akustisch)

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

Konversationsarten

(oral-akustisch)

- Dialog
- Monolog
- Kooperativer Monolog
- Rede

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

P-Wert < 0.001

↠ oral-akustische Texte nutzen mehr Diskursmarker als literat-schriftliche

2. Textpositionen

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

	Anfang	Mitte	Ende
Diskursarten	oral-akustische	literat-schriftliche	oral-akustische
Genres	interaktionslastige	interaktionslastige	sachliche
Konversationsarten	improvisierte interaktive		geskriptete passive

3. Satzpositionen

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

	Anfang	Mitte	Ende
Diskursarten	oral-akustische	literat-schriftliche	
Genres	sachliche	interaktionslastige	
Konversationsarten	improvisierte interaktive		geskriptete

4. Diskursmarker-Klassen

oral-akustische Texte nutzen mehr

COMPARISON und CONTINGENCY

Diskursmarker als literat-schriftliche

5. Häufigste Diskursmarker der Klassen

as	TEMPORAL
and	EXPANSION
but	COMPARISON
for	CONTINGENCY

3 häufigste Diskursmarker:

> 80% aller Contingency Marker

häufiger in oral-akustisch:

because, now that, whatever

häufiger in literat-schriftlich:

in response to, since

- umgangssprachlichere Diskursmarker eher in improvisierten und oral-akustischen Texten
- gehobenere Diskursmarker eher in literat-schriftlichen und 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

Alle Beispielsätze wurden entnommen aus *The Lord of the Rings* von J.R.R. Tolkien, veröffentlich bei HarperCollins*Publishers*, 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/10.1080/03637755509375133

[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 dierent discourse types: Literate vs. oral*, 2019. https://www.aclweb.org/anthology/W19-1407/