

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

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

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

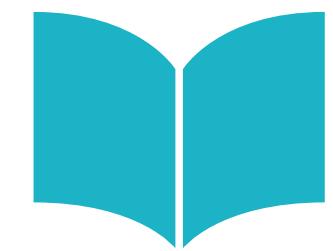
MOTIVATION

- 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



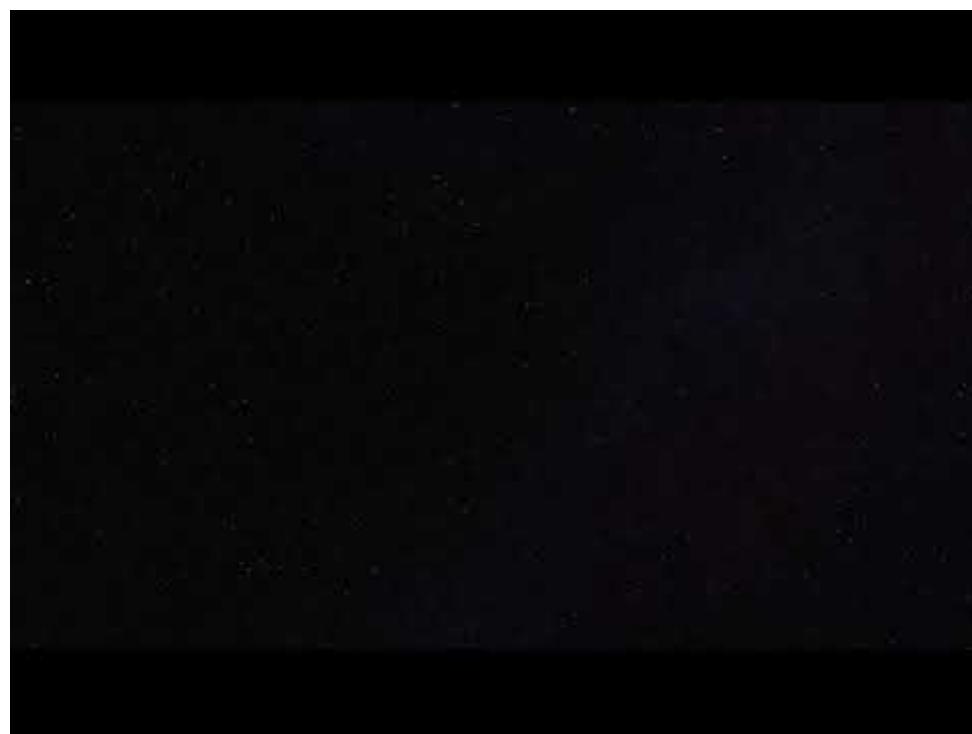
literates Konzept – schriftliches Medium



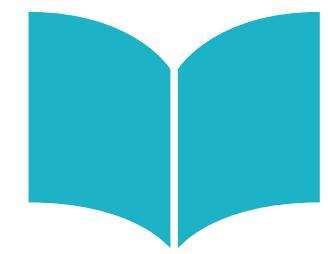
MOTIVATION

- 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



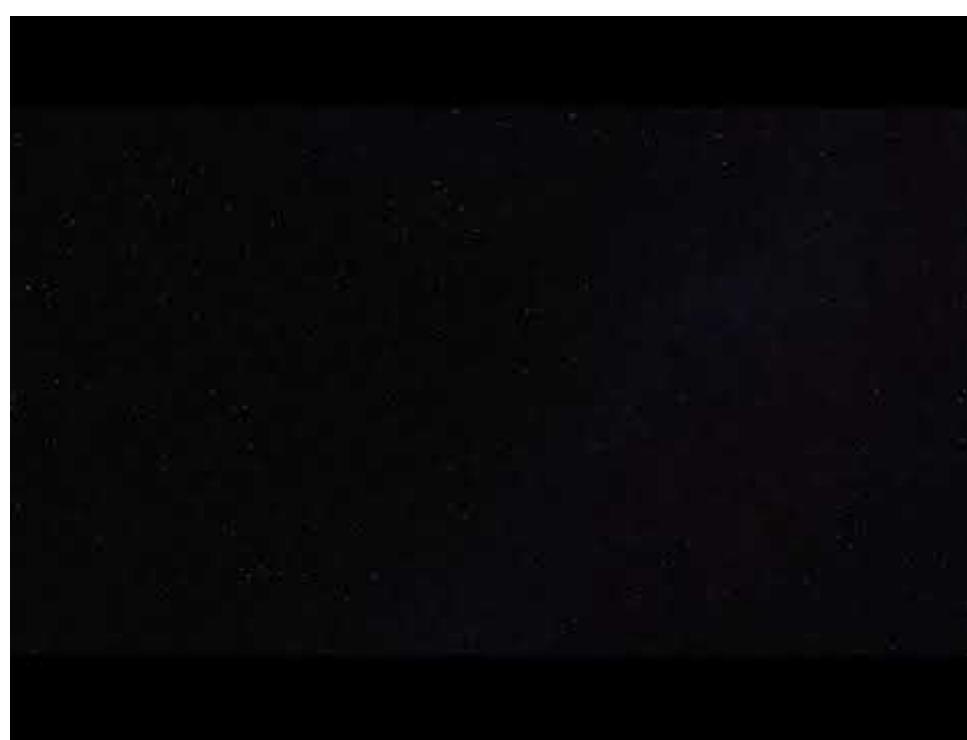
literates Konzept – schriftliches Medium



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

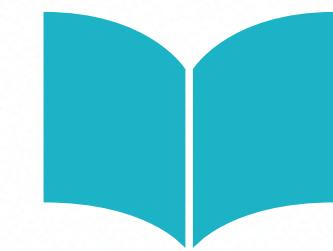


literates Konzept – schriftliches Medium



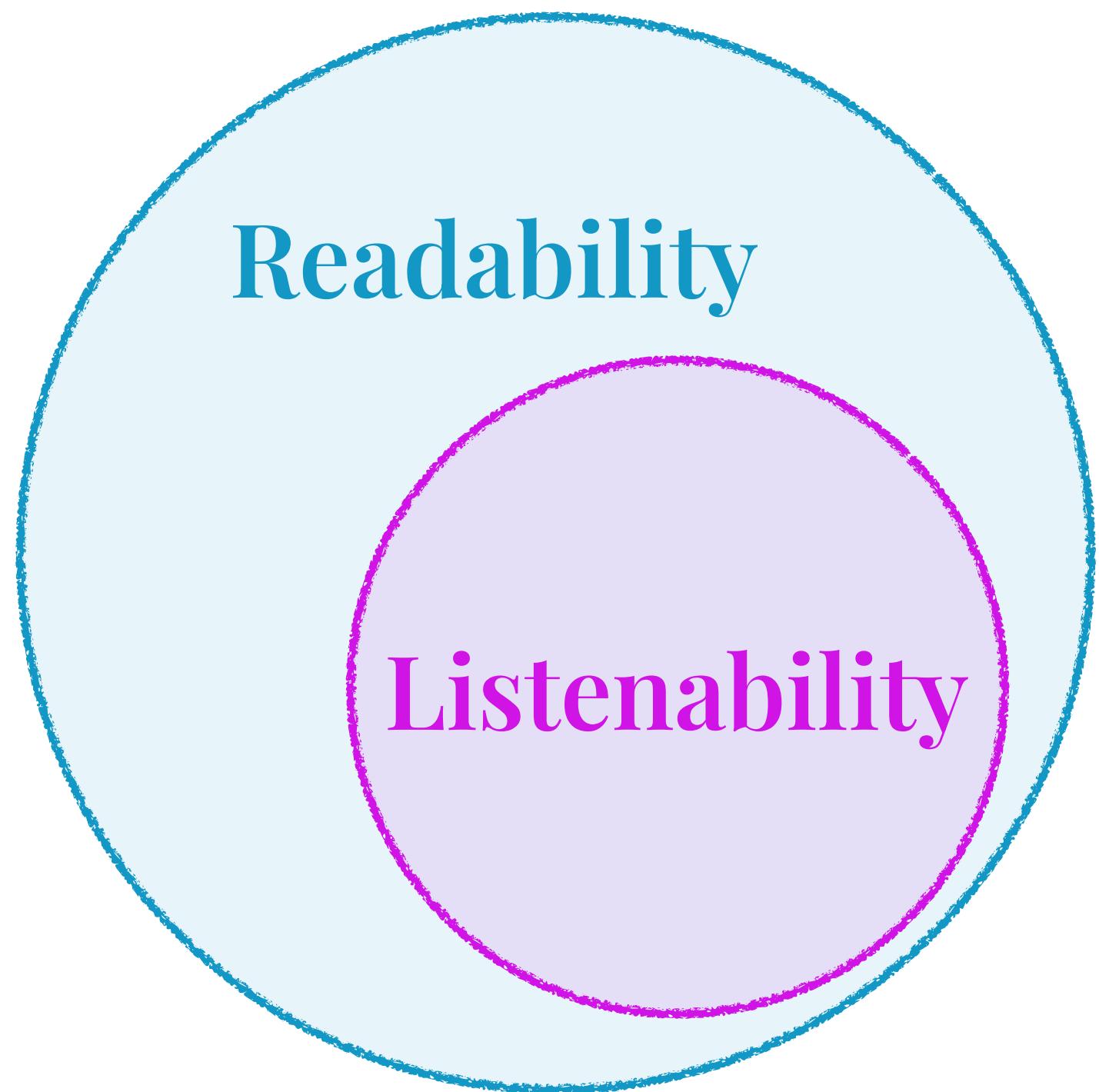
By Dennis Overbye

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.



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

mündliche Tradition



Quelle



schriftliche Tradition



Quelle

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

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

DISKURSMARKER

Begriff

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

- lexikaler Ausdruck, kann nicht flektiert werden
- signalisiert zweiseitige Relation zwischen Diskurssegmenten

DISKURSMARKER

Begriff

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

- lexikaler Ausdruck, kann nicht flektiert werden
- signalisiert zweiseitige Relation zwischen Diskurssegmenten

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

DISKURSMARKER

Begriff

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

- lexikaler Ausdruck, kann nicht flektiert werden
- signalisiert zweiseitige Relation zwischen Diskurssegmenten

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

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*)

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*)

DISKURSMARKER

Zusammenfassung

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

DISKURSMARKER

Bedeutungsgruppen

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

DISKURSMARKER

Bedeutungsgruppen

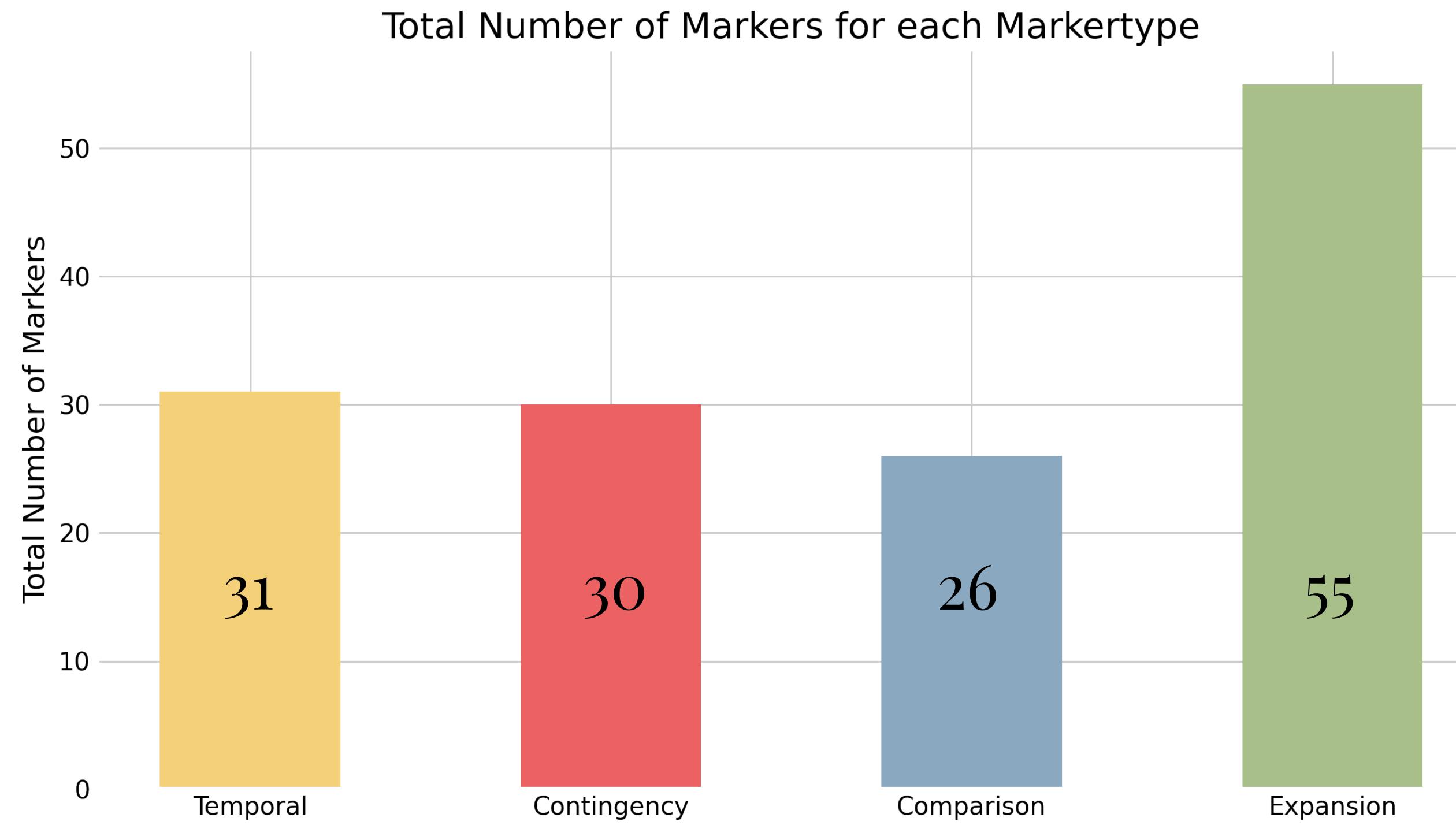
Nach Das et al., 2018 [1]
(EnDimLex)

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, ...



*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:

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)

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)

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)

TEXTDATEN

Corpora

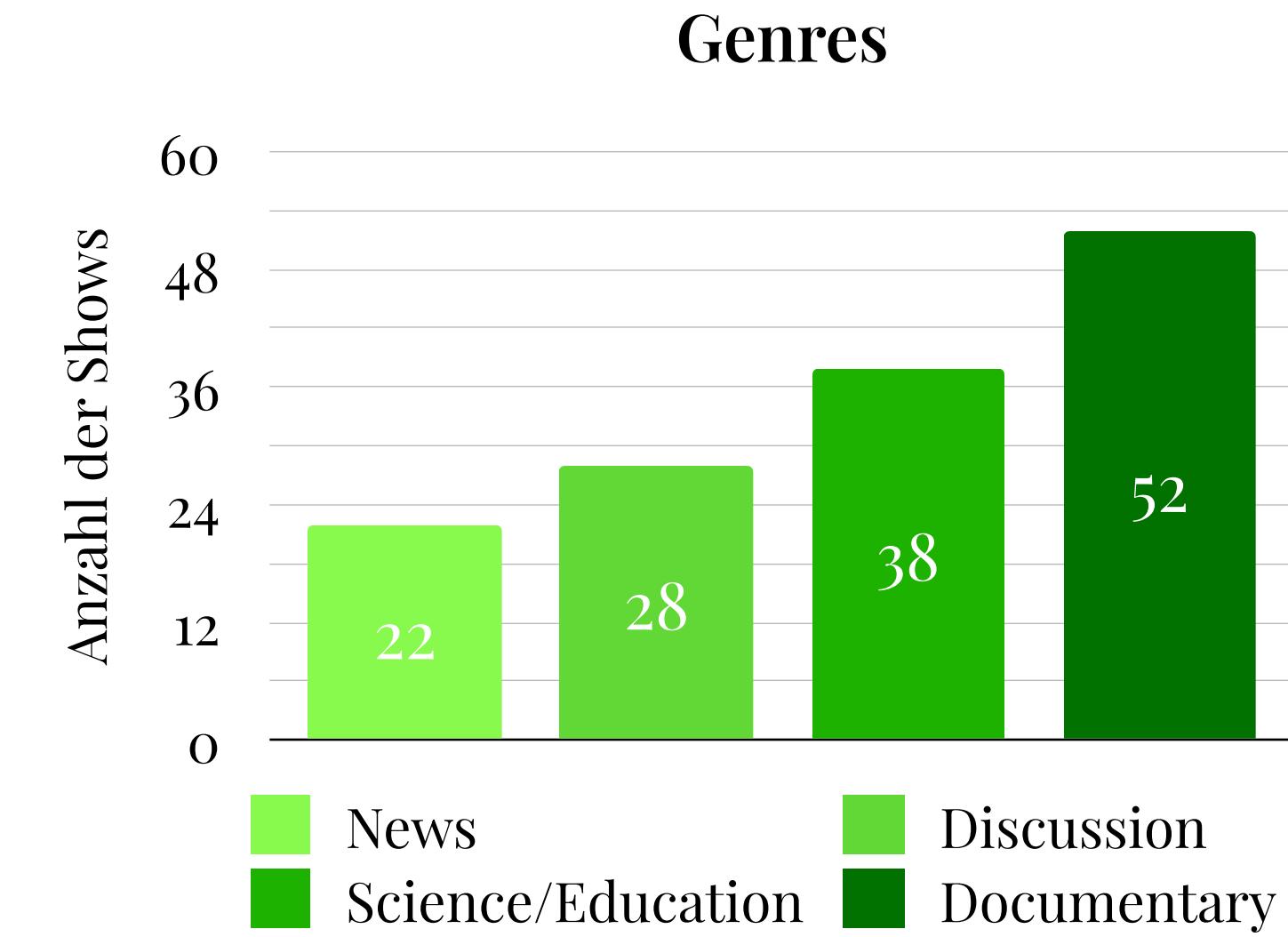
schriftliche Corpora mit ursprünglich schriftlichem Material &
akustische Corpora mit Transkripten von Audiomaterial

KRITERIEN

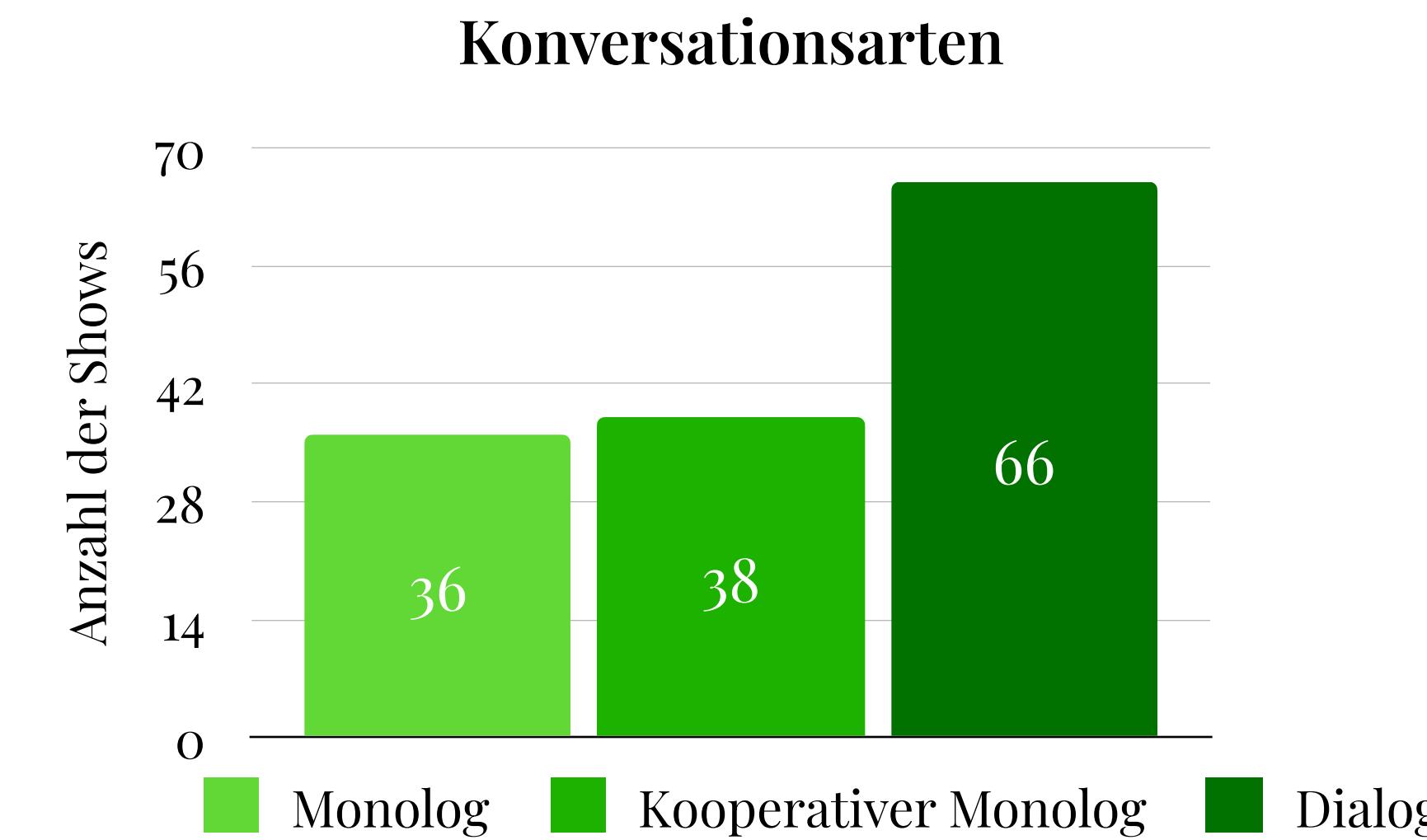
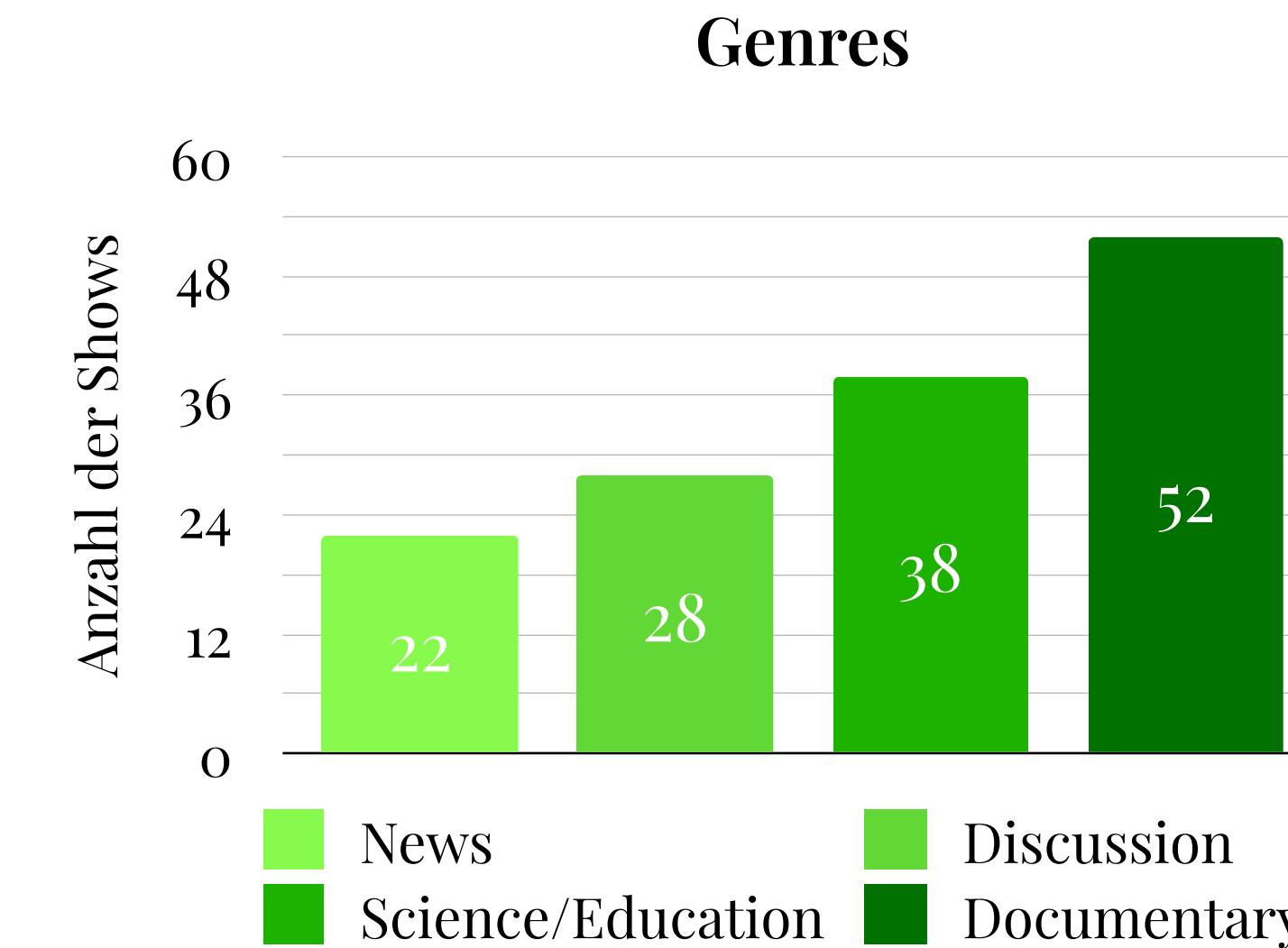
- kostenlos
- groß
- qualitativ hochwertig
- nachrichtenähnlich

CORPUS	Spotify Podcast Corpus
DATEN	<ul style="list-style-type: none">• fast 60.000 Stunden transkribiertes Audiomaterial• verschiedene Produzenten• WER: 18,1 %
RELVANT	<ul style="list-style-type: none">• 140 Shows, 2.782 Episoden• ca. 17 Mio. Wörter

CORPUS	Spotify Podcast Corpus
DATEN	<ul style="list-style-type: none"> • fast 60.000 Stunden transkribiertes Audiomaterial • verschiedene Produzenten • WER: 18,1 %
RELVANT	<ul style="list-style-type: none"> • 140 Shows, 2.782 Episoden • ca. 17 Mio. Wörter



CORPUS	Spotify Podcast Corpus
DATEN	<ul style="list-style-type: none"> • fast 60.000 Stunden transkribiertes Audiomaterial • verschiedene Produzenten • WER: 18,1 %
RELVANT	<ul style="list-style-type: none"> • 140 Shows, 2.782 Episoden • ca. 17 Mio. Wörter



CORPUS	TED-LIUM 3 Corpus
DATEN	<ul style="list-style-type: none">• 1.983 TED-Talks• ca. 4 Mio. Wörter• WER: 6,7 %
RELEVANT	<ul style="list-style-type: none">• Alle Talks

CORPUS	TED-LIUM 3 Corpus
DATEN	<ul style="list-style-type: none">• 1.983 TED-Talks• ca. 4 Mio. Wörter• WER: 6,7 %
RELEVANT	<ul style="list-style-type: none">• Alle Talks

Genre



Konversationsart



CORPUS	New York Times Corpus
DATEN	<ul style="list-style-type: none">• 1,8 Mio. Nachrichtenartikel der New York Times• ca. 1,1 Mrd. Wörter
RELEVANT	<ul style="list-style-type: none">• Alles

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

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

TEXTSORTEN

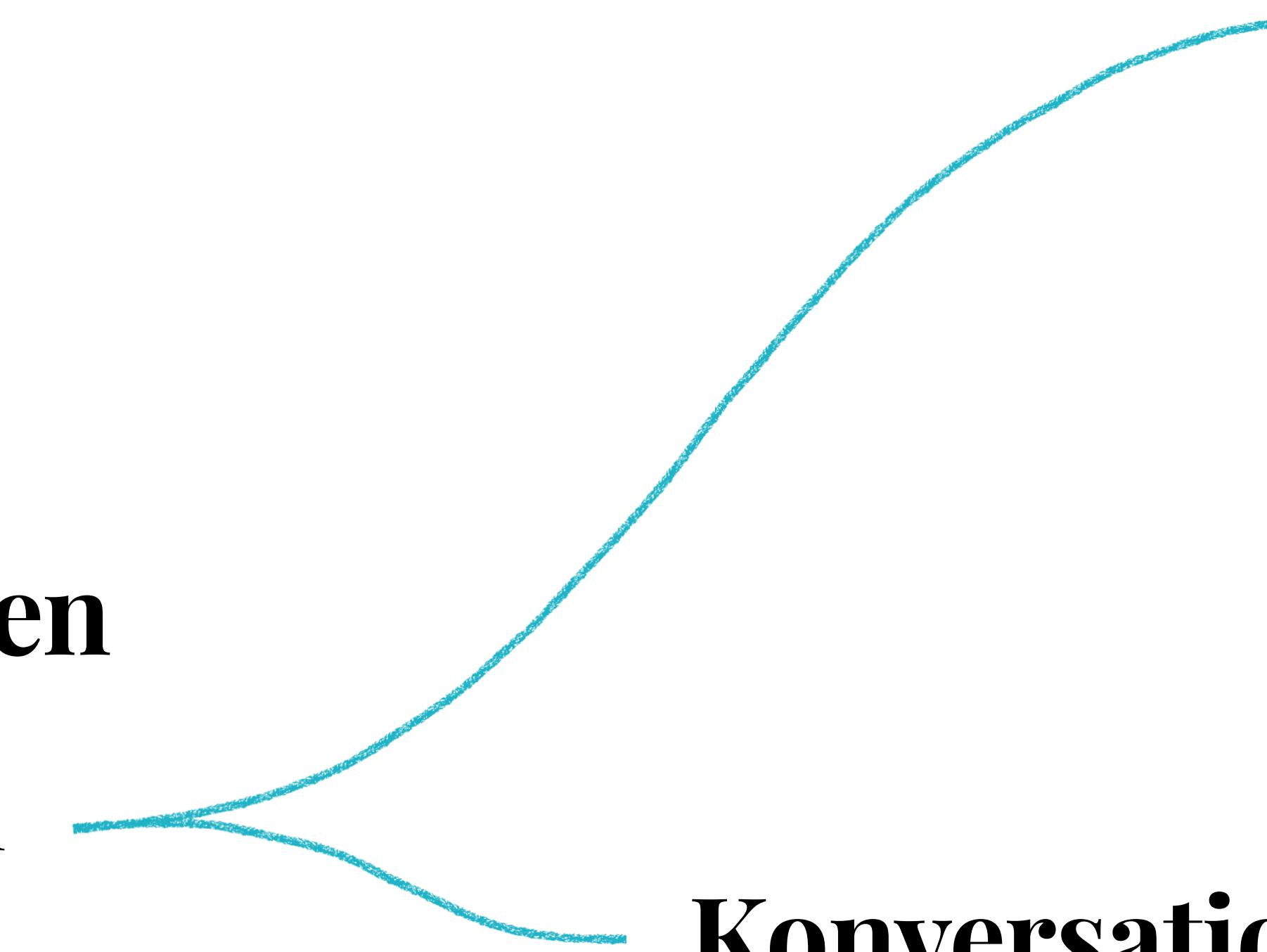
Diskursarten

- oral-akustisch
- literat-schriftlich

TEXTSORTEN

Diskursarten

- oral-akustisch
- literat-schriftlich



Konversationsarten

(oral-akustisch)

- News
- Discussion
- Science/Education
- Documentary
- Presentation
- Dialog
- Monolog
- Kooperativer Monolog
- Rede

AUSWERTUNG

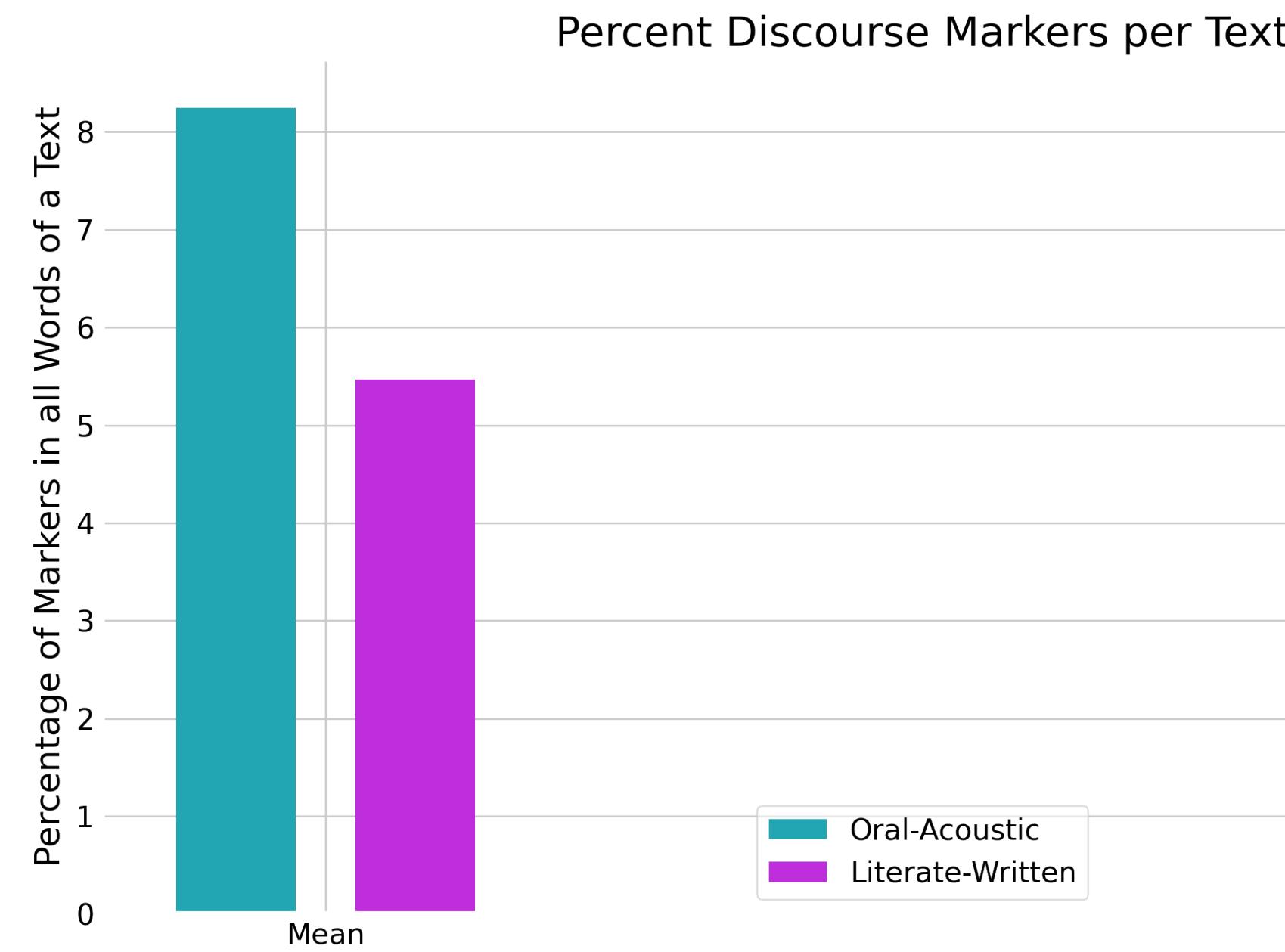
Generelle Verteilung

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

AUSWERTUNG

Generelle Verteilung

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



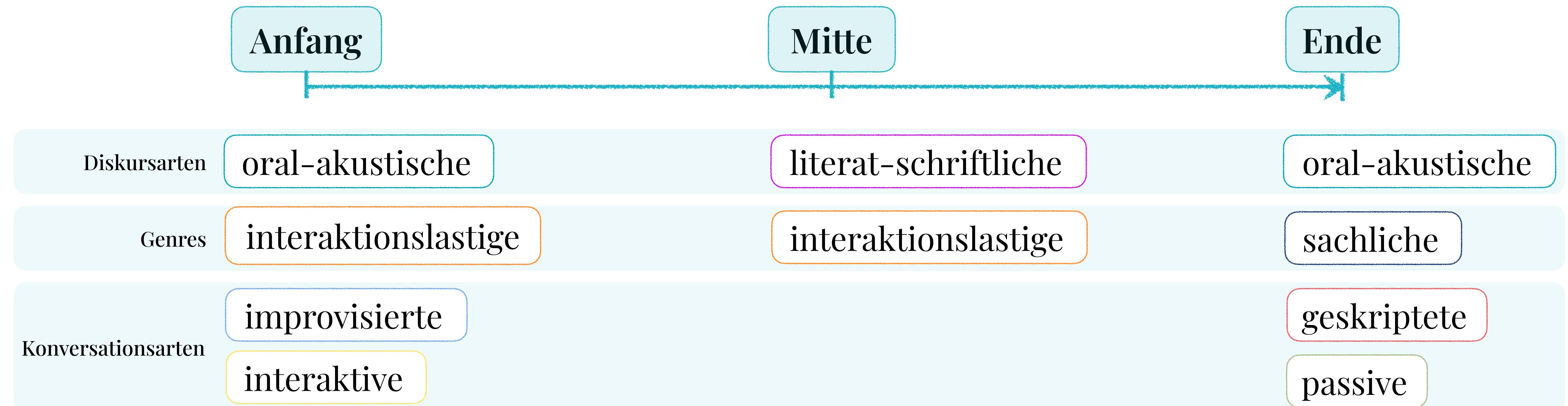
P-Wert < 0.001

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

AUSWERTUNG

Positionen im Text

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



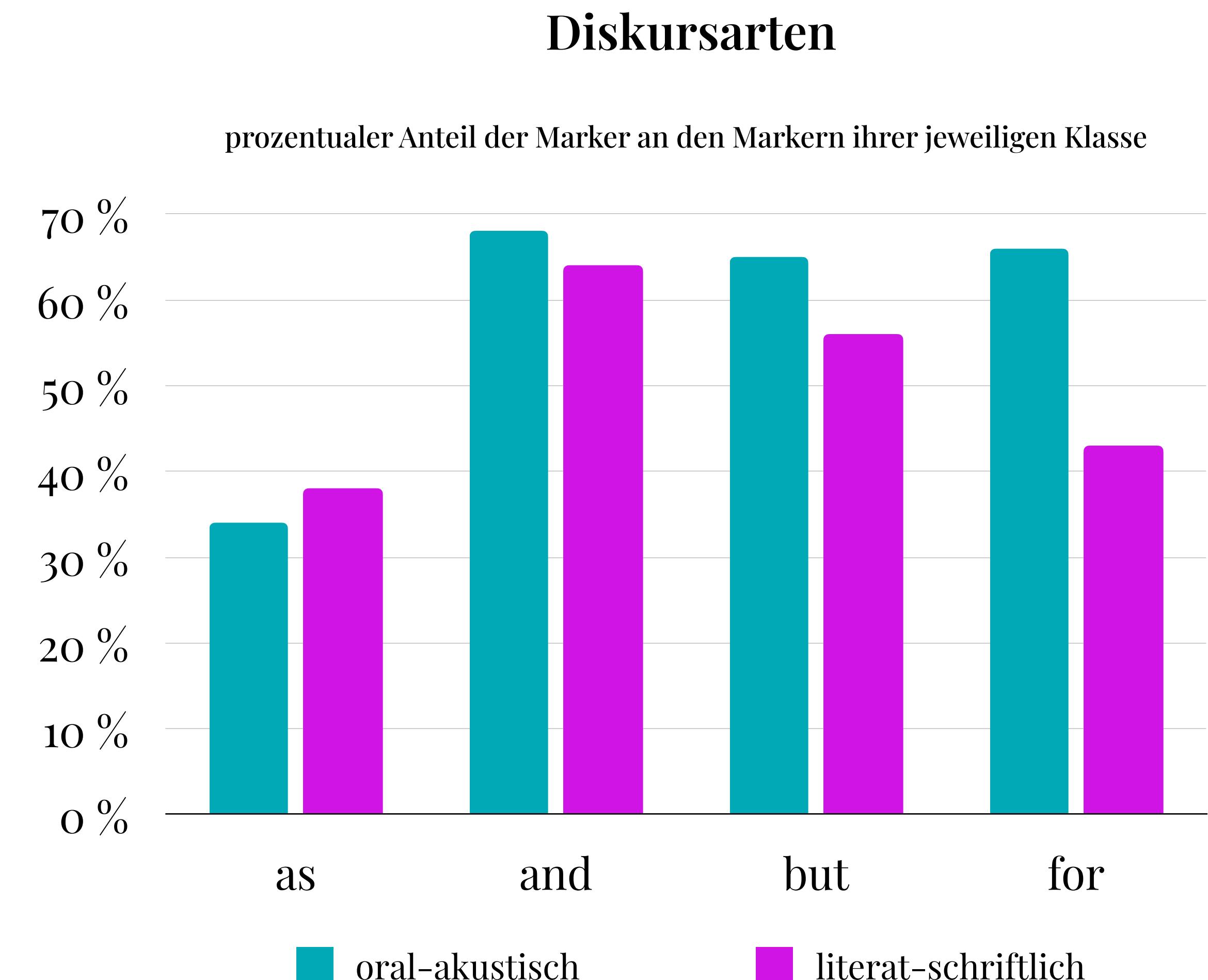
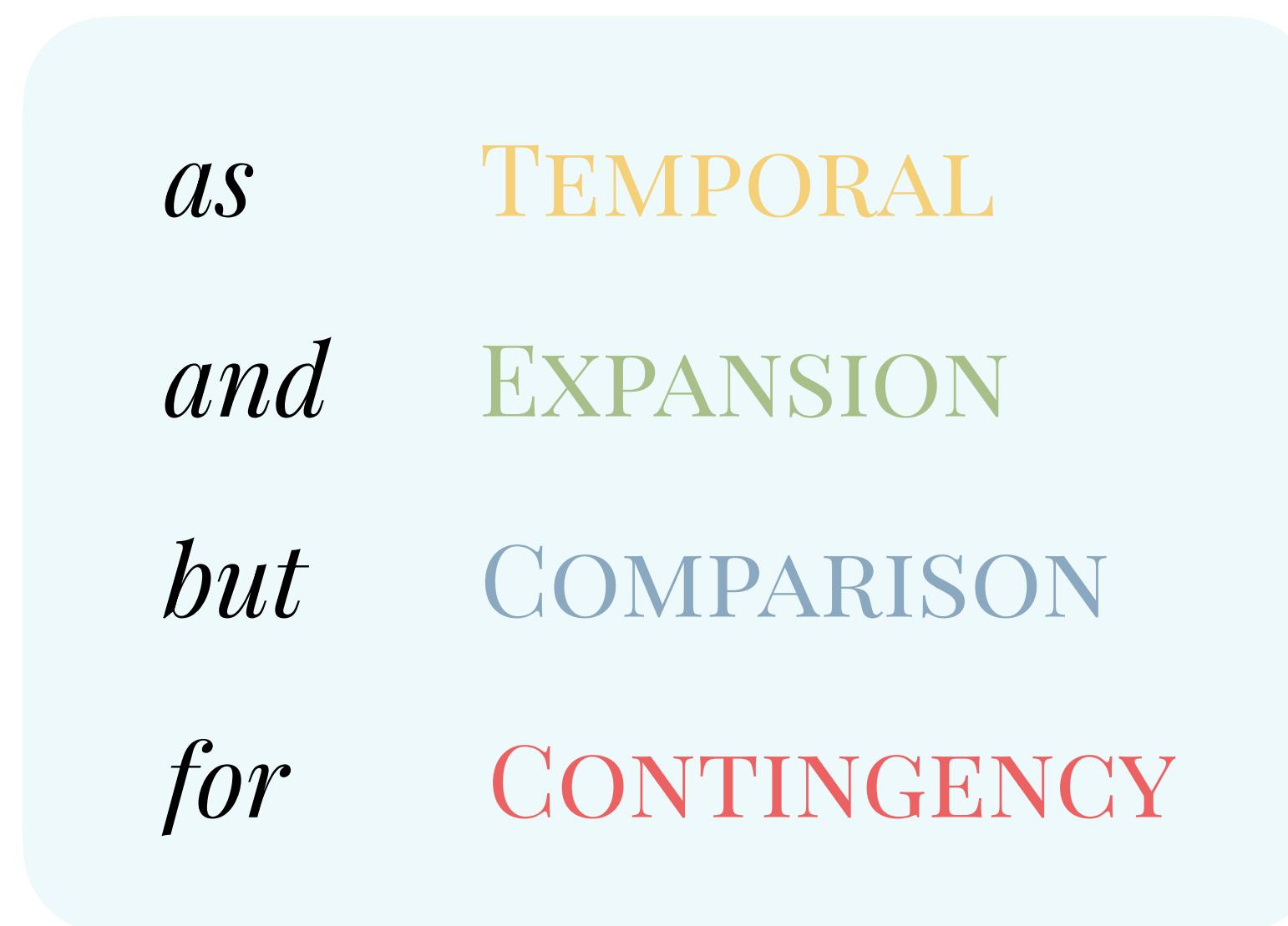
AUSWERTUNG

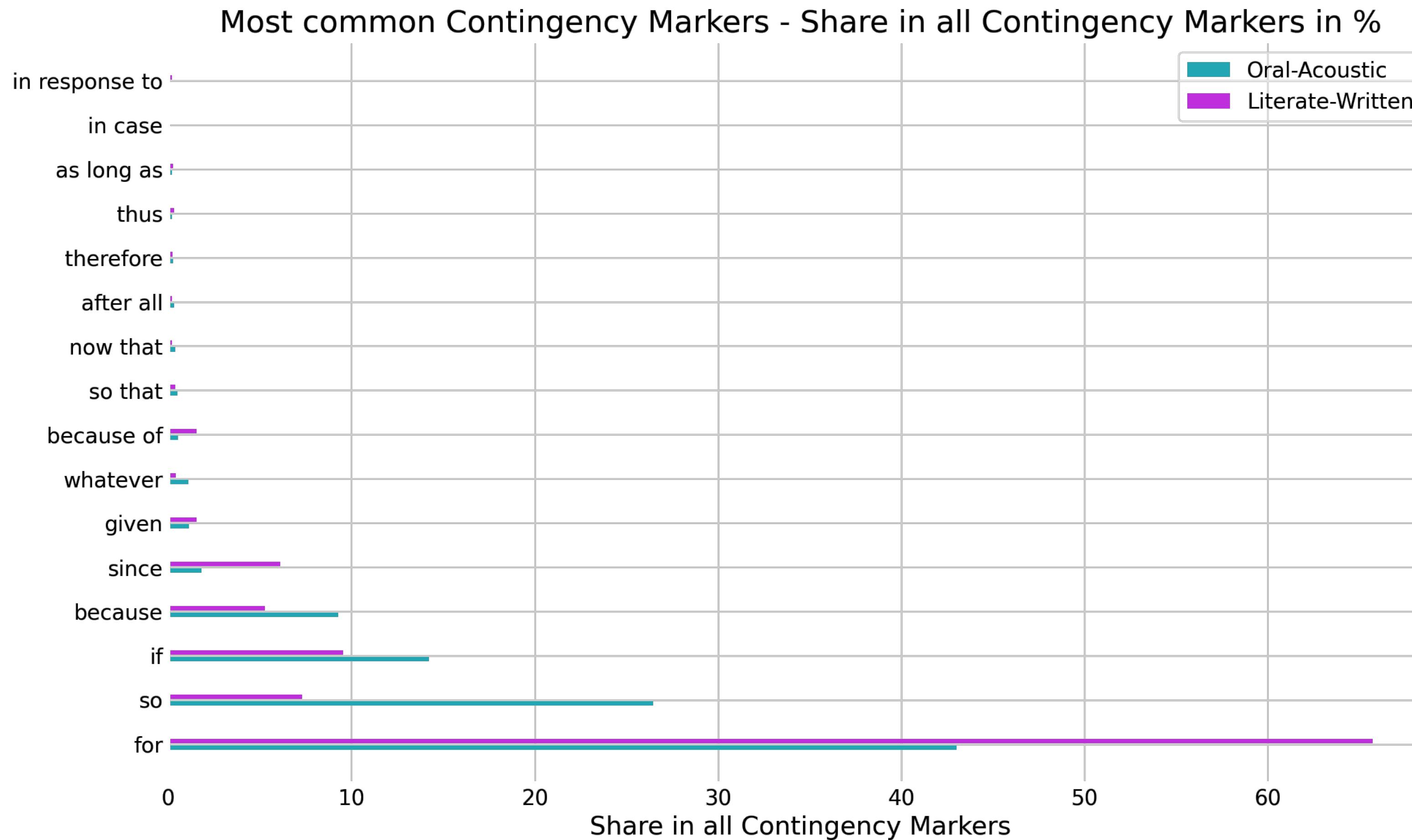
Häufigste Diskursmarker der Klassen

<i>as</i>	TEMPORAL
<i>and</i>	EXPANSION
<i>but</i>	COMPARISON
<i>for</i>	CONTINGENCY

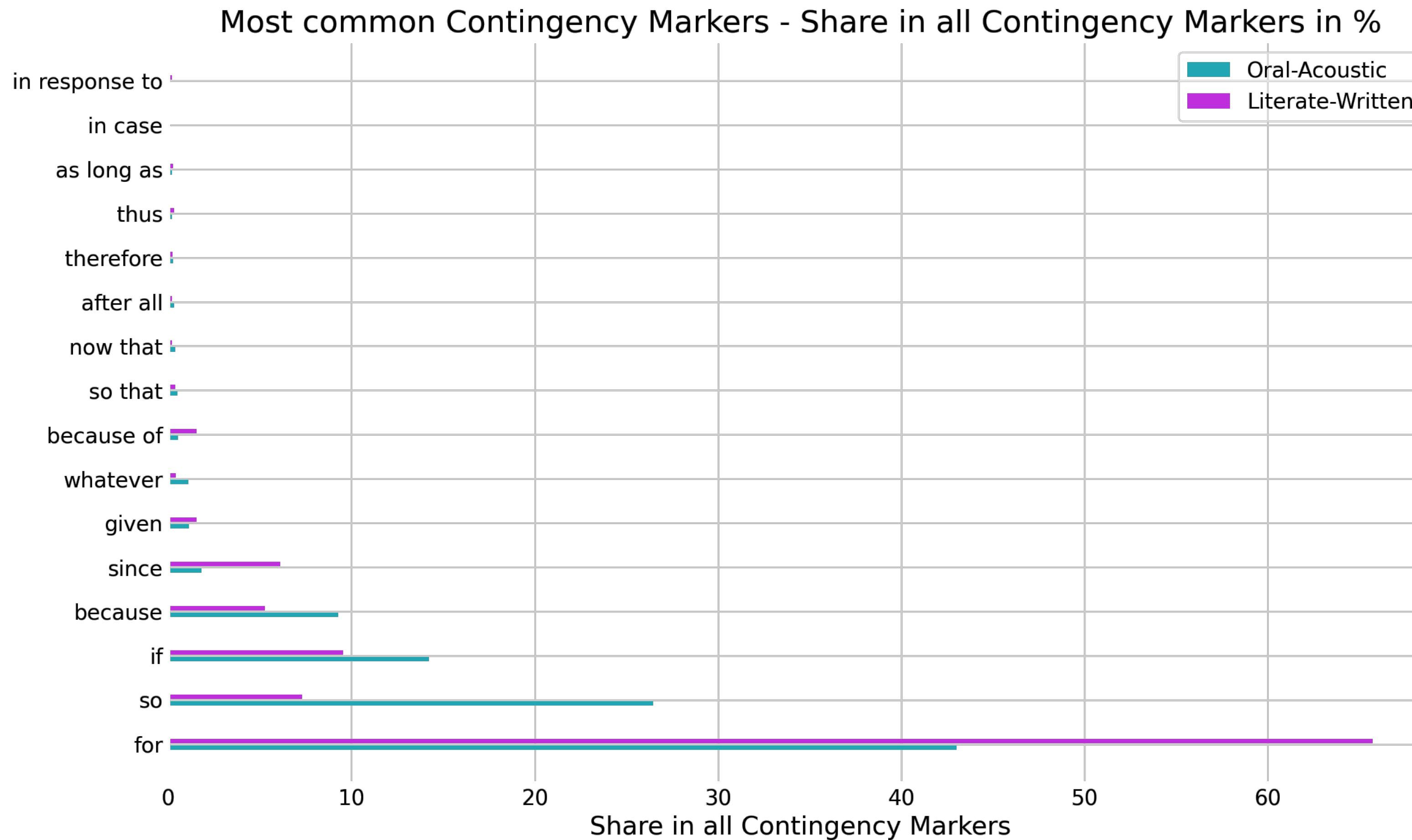
AUSWERTUNG

Häufigste Diskursmarker der Klassen





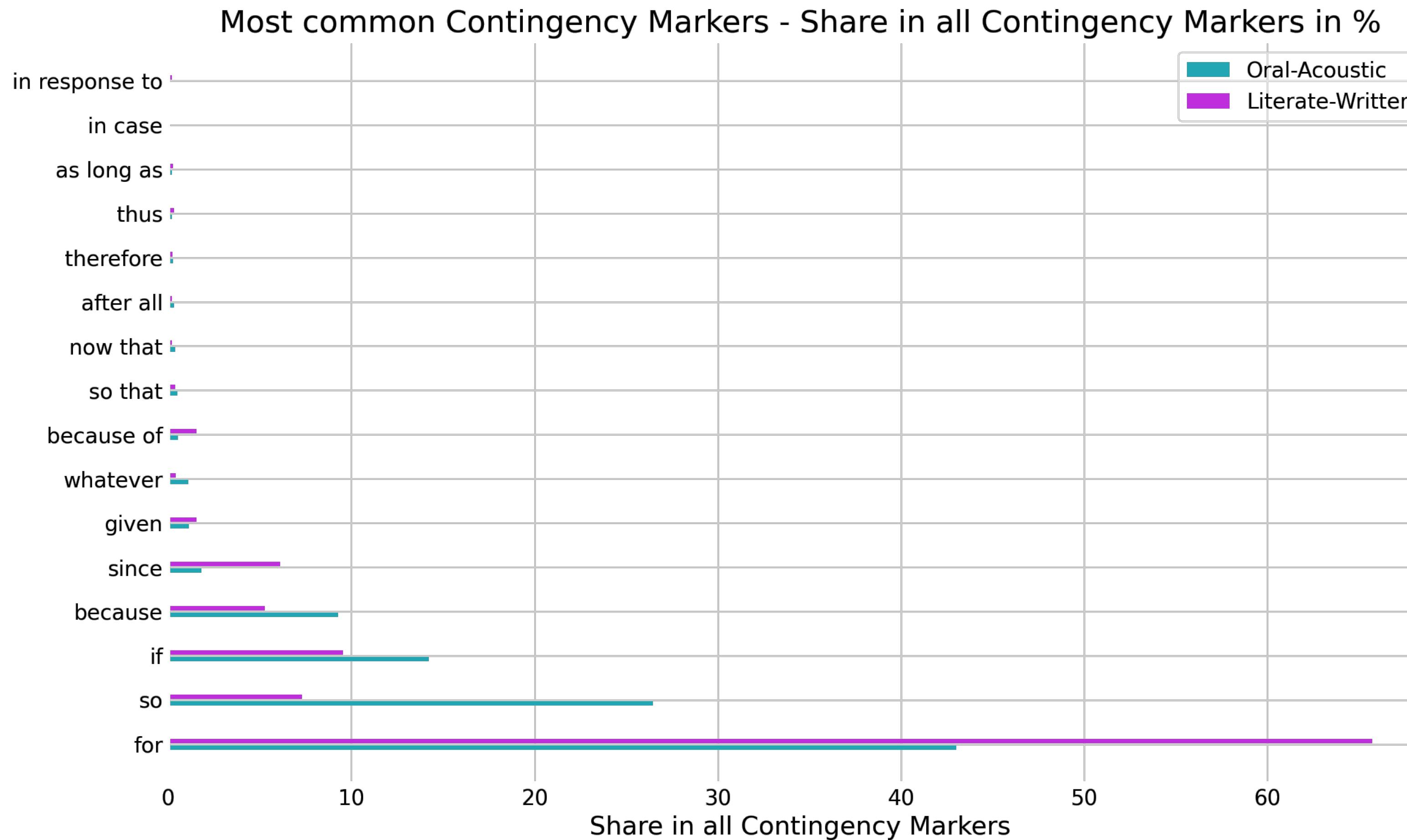
3 häufigste Diskursmarker:
 > 80% aller Contingency Marker



3 häufigste Diskursmarker:
 > 80% aller Contingency Marker

häufiger in oral-akustisch:
because, now that, whatever

häufiger in literat-schriftlich:
in response to, since



- 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

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, 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/>