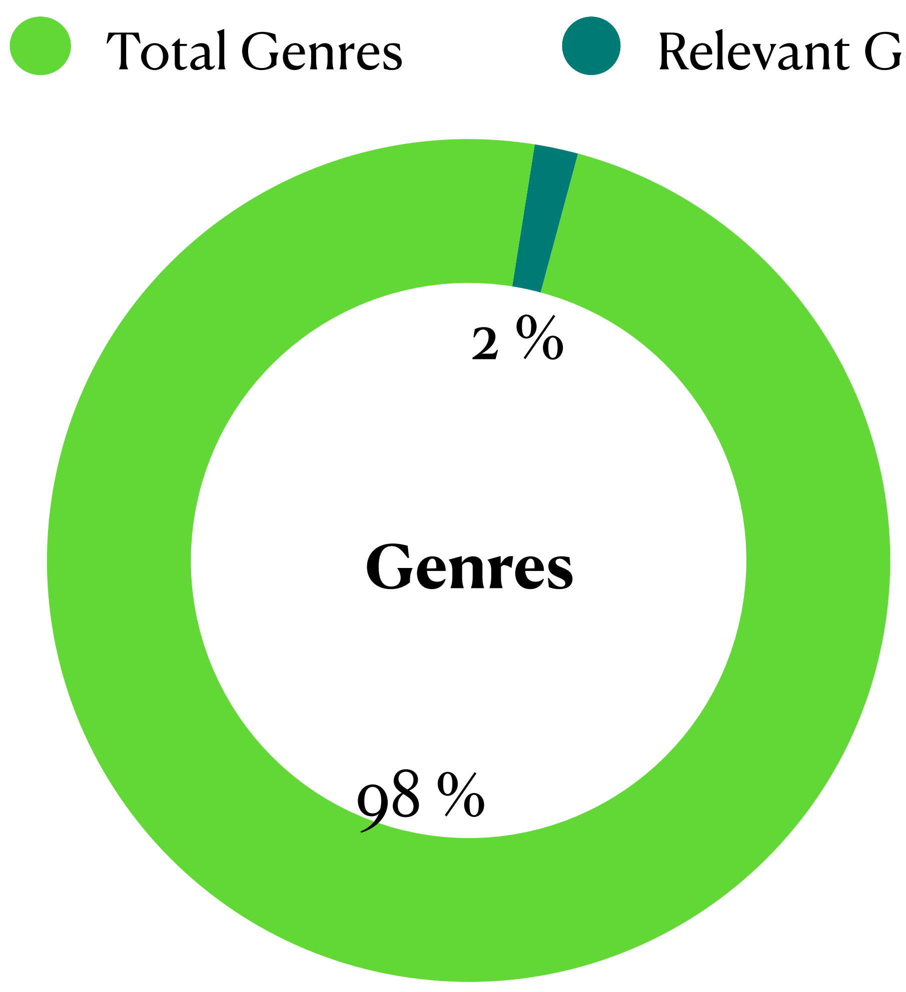
Spotify Podcast Dataset

Spotify Podcast Dataset

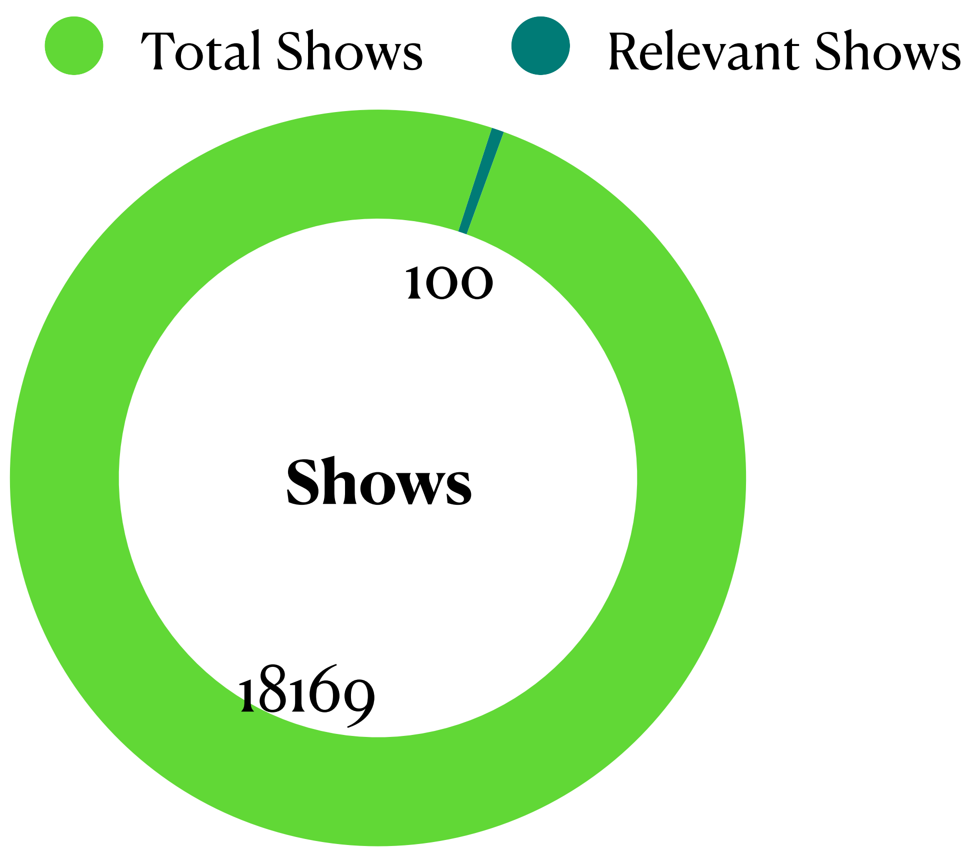
- 100.000 podcast episodes with aligned ASP transcripts
- more than 47.000 hours of transcribed audio
- automatic transcripts
- Word Error Rate: 18.1%
- has punctuation

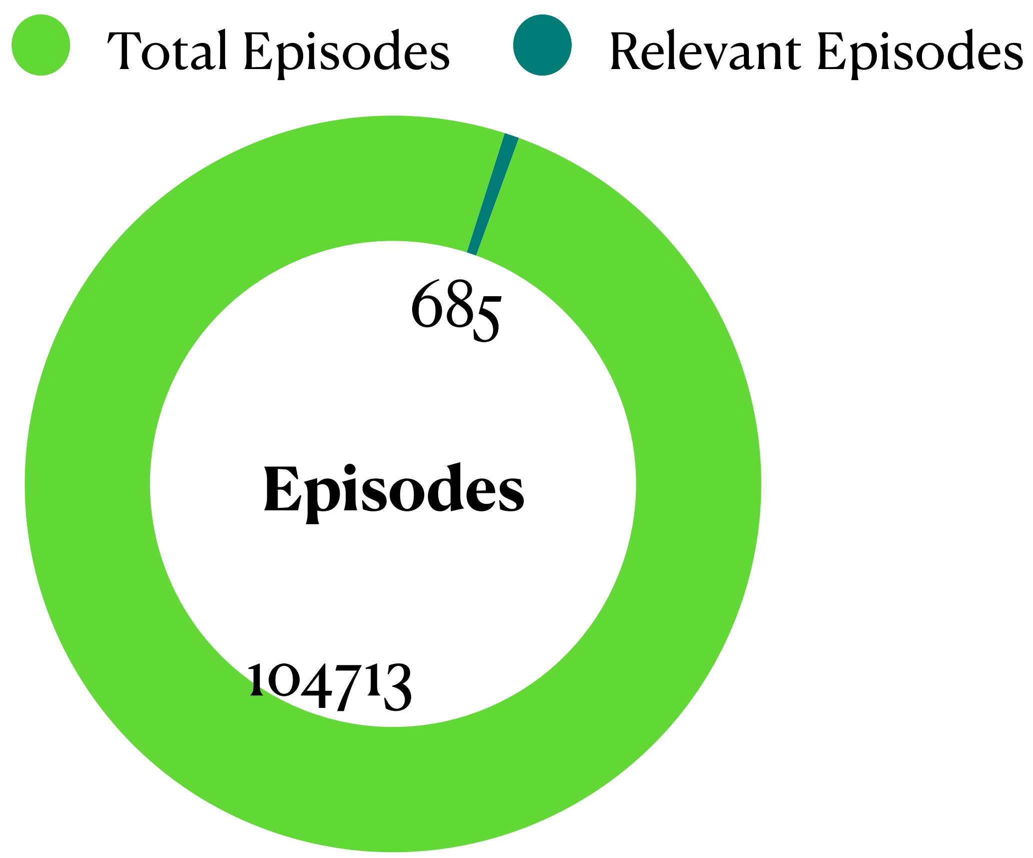
General



Relevant Genres

- Business News
- Daily News
- News
- news
- Sports News
- Tech News



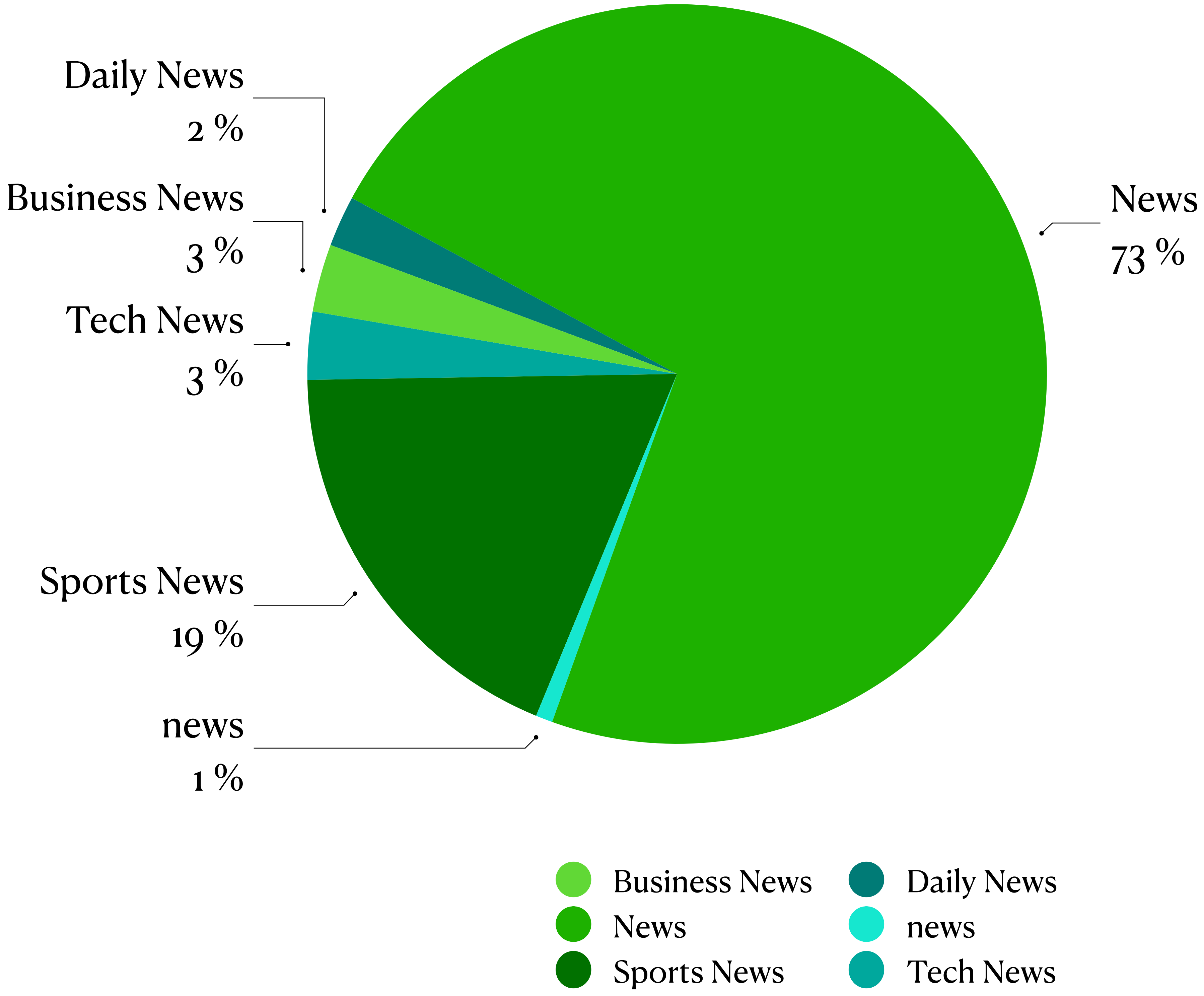


Episode data

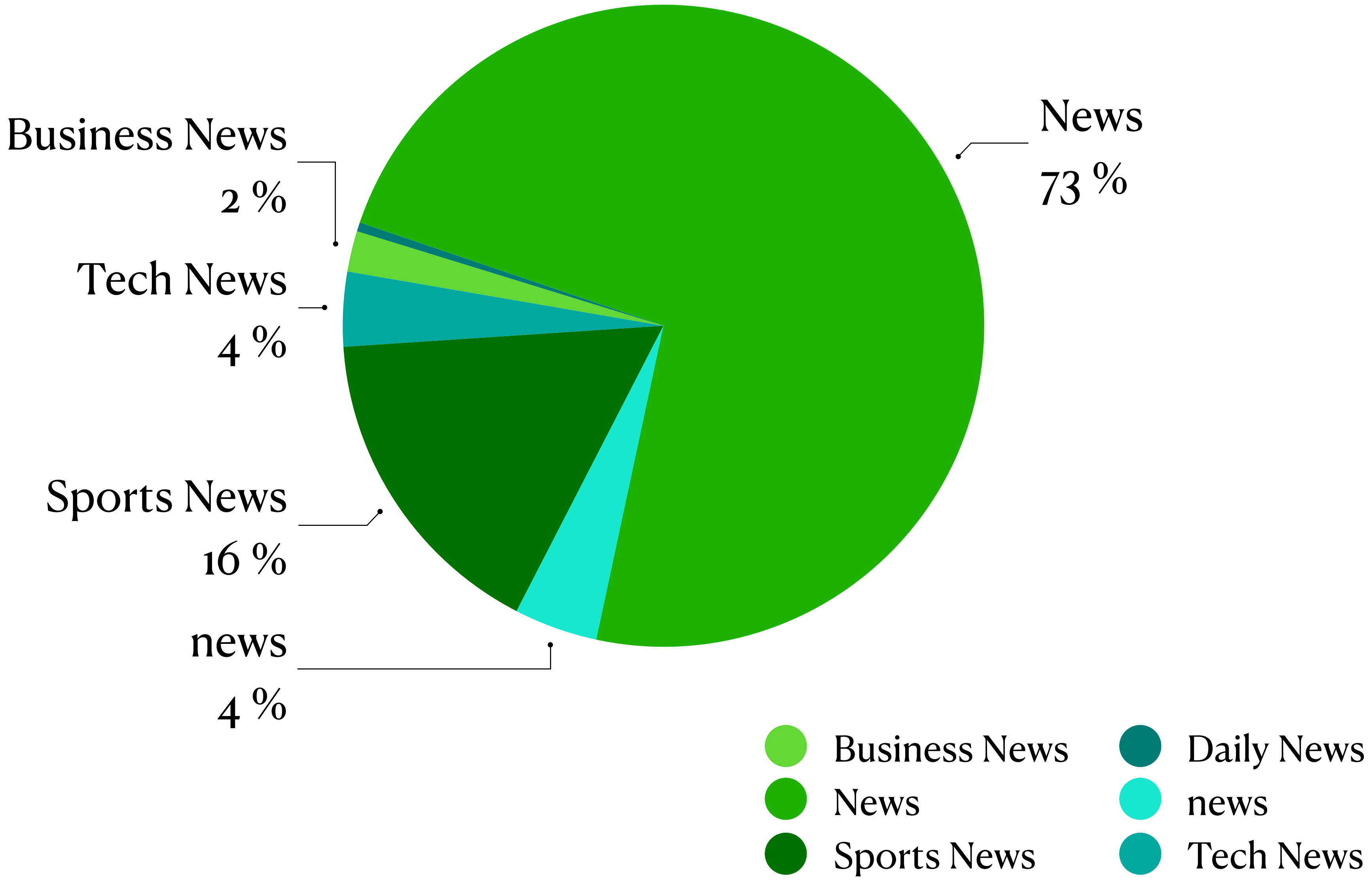
	min	average	max
minutes	1	31,6	305
words	11	5728	43504

Quelle: The Spotify Dataset Paper

Relevant Data



Relevant Shows



Relevant Episodes

* The List of relevant shows contains shows that may have multiple relevant genres and are therefore listet multiple times.

Spotify Podcast Dataset

- We can use approx. 685 Episodes from 100 news-related shows
- The already collected Data contains a list of all relevant shows incl. their respective show_uri
- the transcription files are named after the show_uri

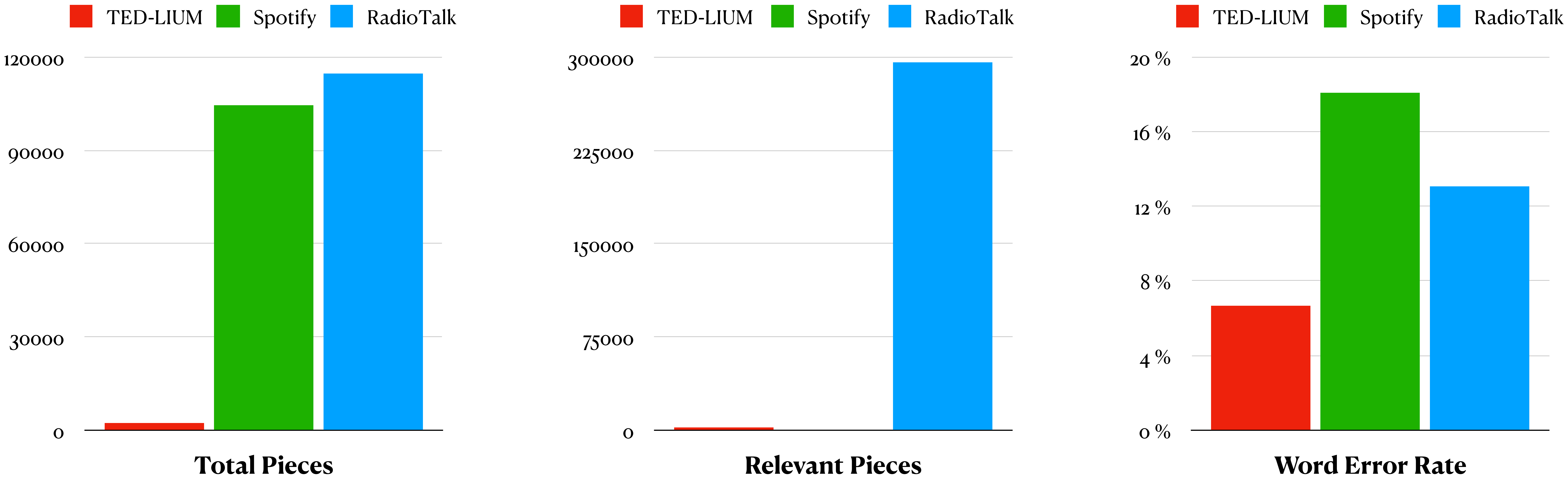
TED-LIUM 3

TED-LIUM 3

- 452 hours of transcribed speech
- 2351 speeches
- 4.9 M words
- automatic transcripts
- Word Error Rate: 6.7%
- no punctuation
- no genre information

Comparison

Comparison



🟥 TED-LIUM
🟩 Spotify
🟦 RadioTalk
120000
90000
60000
30000
0
Total Pieces
300000
225000
150000
75000
0
Relevant Pieces
20 %
16 %
12 %
8 %
4 %
0 %
Word Error Rate

Comparison

Checklist

Category	TED-LIUM	Spotify	RadioTalk
Punctuation	☐	☑	☐
Genre-Info	☐	☑	☐
Contains News	☐	☑	☑
Contains Opinion Pieces	☑	☑	☑
„Good enough“ Transcripts	☑	☑	☐
Enough relevant Data	☑ ★	☑	☑
Human-made Transcripts	☐	☐	☐
RSS-Feed	☐	☑	☐
Free	☑	☑	☑
Variety of Producers	☑	☑	☑

★ according to Ortmann and Dipper (2019) speeches have the same *index of orality* as the texts we want to produce:

- many participants = -1
- monologue = -1
- synchronous production = -1
- asynchronous reception = 1

—> **Index of orality** = **-2**