Corpora Analysis

For the usage in conversational News

Question

Which data can we use?

- data from news (related) shows
- data from scripted programs, not much improvising (-> writing for listening)

Criteria

- ☐ Free
- ☐ Variety of Producers
- ☐ Punctuation
- ☐ Genre-Info
- ☐ Contains News
- ☐ Contains Opinion Pieces
- ☐ „Good enough“ Transcripts
- ☐ Enough relevant data
- ☐ Human-made Transcripts
- ☐ RSS-Feed

Word Error Rate

Rates the accuracy of transcripts created by speech recognition APIs

$$WER = \frac{S + I + D}{N}$$

Substitution - number of replaced words

Insertions - number of inserted words

Deletions - number of omitted words

N - number of words in the reference/original (what has really been said)

Word Error Rate

Some discussion whether a lower error rate really indicates a better speech recognition:

- all words are „worth“ the same
- no different scores for e.g. „It’s a matter of free peach“ and „It’s a matter of free“
- noisy data is a problem
- normalization discrepancies (punctuation, speaker turns, glue words, phonetic reductions, ...)
- dependent on the reference / size of reference data

Measures

Word Error Rate

TED-LIUM

- Kaldi toolkit
- WER 6.7%
- no information about how this number was obtained

Spotify Podcasts

- Google Cloud Platform's Speech-to-Text API (GCP-ASR)
- Manual evaluation of transcript-quality on 1600 Episodes
- WER 18.1 %

RadioTalk

- Kaldi toolkit
- WER 13.1 %
- „Measured on a set of human transcribed talk radio content that aired after the time period of the system's training data"

(quote from the RadioTalk Paper)

Index of Orality

- Index that shows the orality of a text, high scores meaning orally-oriented, introduced by Ortmann and Dipper (2019)
- They assigned an index of orality to each register of data they referred to (News, Speech, TED, Chat, Dialogue), based on 4 factors:
  - Participants (many = -1, few = 1)
  - Interactiveness (monologue = -1, dialogue = 1)
  - Production Circumstances (synchronous = 1, quasi-synchr. = 0, asynchronous = -1)
  - Reception Circumstances (synchronous = 1, quasi-synchr. = 0, asynchronous = -1)

Index of Orality

Register	Participants		Interactiveness		Production		Reception		Index of Orality
	value	score	value	score	value	score	value	score	score (sum)
News	many	-1	monolog	-1	asynchronous	-1	asynchronous	-1	-4
Speech	many	-1	monolog	-1	asynchronous	-1	synchronous	1	-2
TED	many	-1	monolog	-1	quasi-synchr.	0	synchronous	1	-1
Chat	few	1	dialog	1	quasi-synchr.	0	quasi-synchr.	0	2
Dialog	few	1	dialog	1	synchronous	1	synchronous	1	4

Table 3: Expected orality based on four situational characteristics of the registers. The characteristics rank the registers from highly literate (*News*) to highly oral (*Dialog*).

Quelle: Variation between different discourse types: literal vs. oral by Katrin Ortmann und Stefanie Dipper (2019)

„Our“ Index of orality:

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

Index of Orality = -2

Conversational News Texts have a relatively low index of orality, comparable to Speeches and TED-Talks

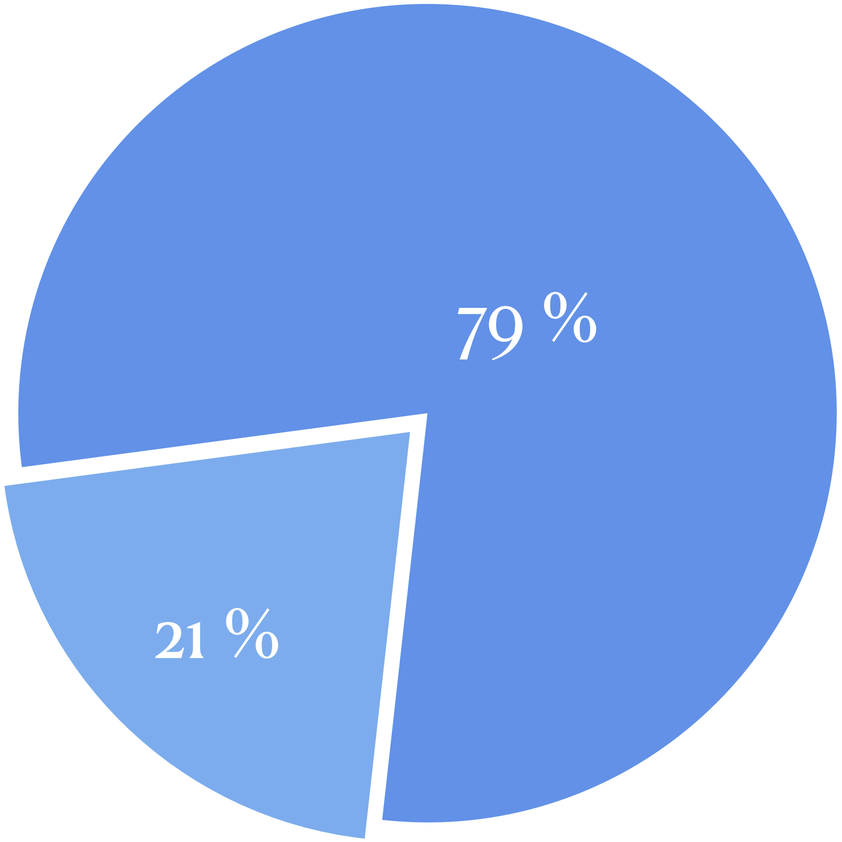
RadioTalk Corpus

- 284 000 hours of data
- 2.8 billion words
- Total Entries: 115.769.559
- Total Shows: 1010
- Word Error Rate: 13.1%
- no punctuation
- genre only identifiable by title

```
{
"content":"The transcribed speech from the snippet",
"callsign": "The call letters of the station the snippet aired on",
"city":"The city the station is based in, as in FCCC filings",
"state":"The state the station is based in, as in FCCC filings",
"show_name":"The name of the show containing this snippet",
"signature":"The initial 8 bytes of an MD5 hash of the content field, after lowercasing and removing English stopwords (specifically the NLTK stopword list), intended to help with deduplication",
"studio_or_telephone":"A flag for whether the underlying audio came from a telephone or studio audio equipment. (The most useful feature in distinguishing these is the narrow frequency range of telephone audio.)",
"guessed_gender":"The imputed speaker gender",
"segment_start_time":"The Unix timestamp of the beginning of the underlying audio",
"segment_end_time":"The Unix timestamp of the end of the underlying audio",
"speaker_id":"A diarization ID for the person speaking in the audio snippet",
"audio_chunk_id":"An ID for the audio chunk this snippet came from (each chunk may be split into multiple snippets)"
}
```

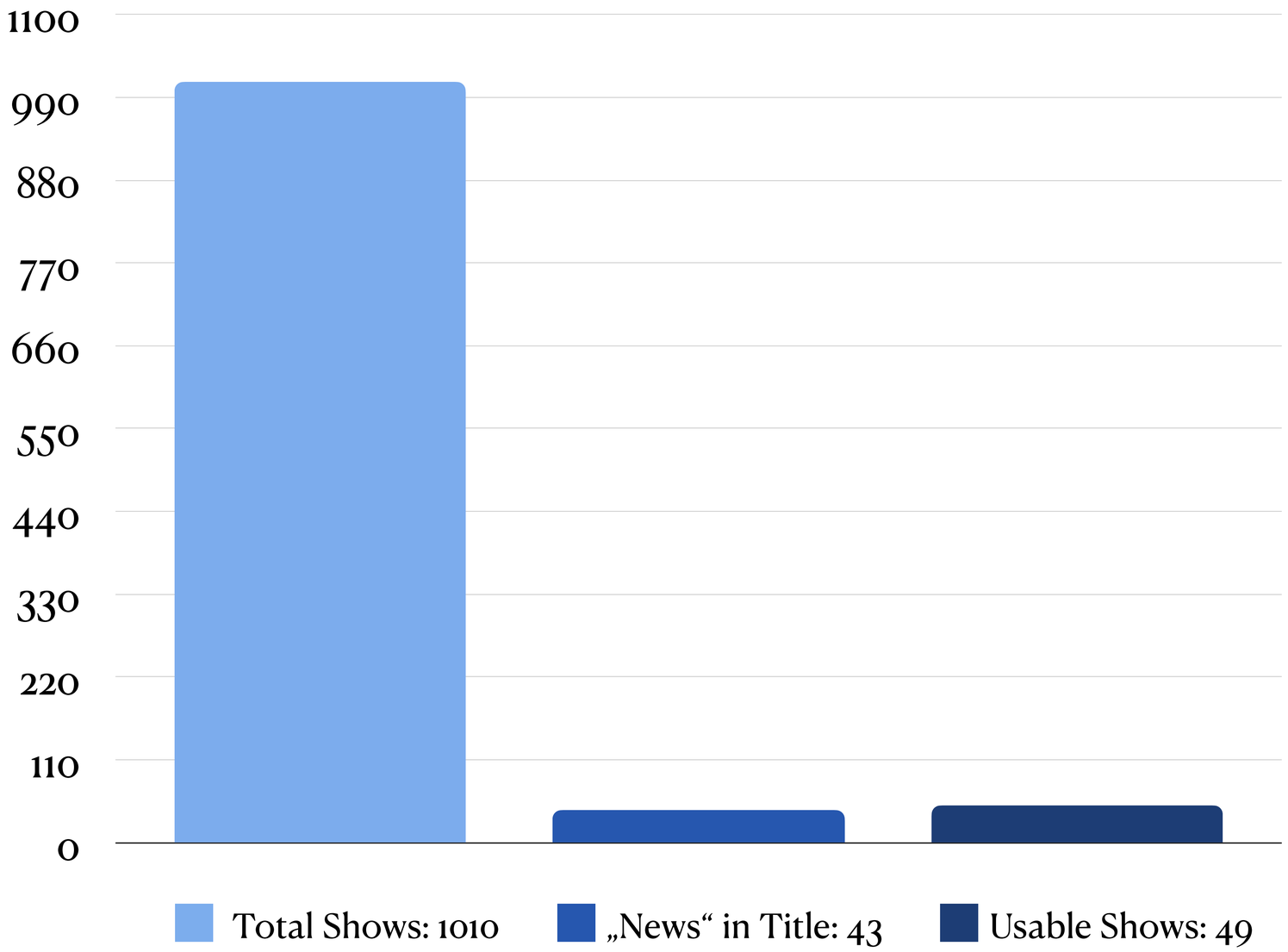
From a total of **115.769.559** snippets ...

- **91.310.616** are tagged with a show name so the genre can be identified
- While **24.458.943** are unidentifiable and therefore useless for us



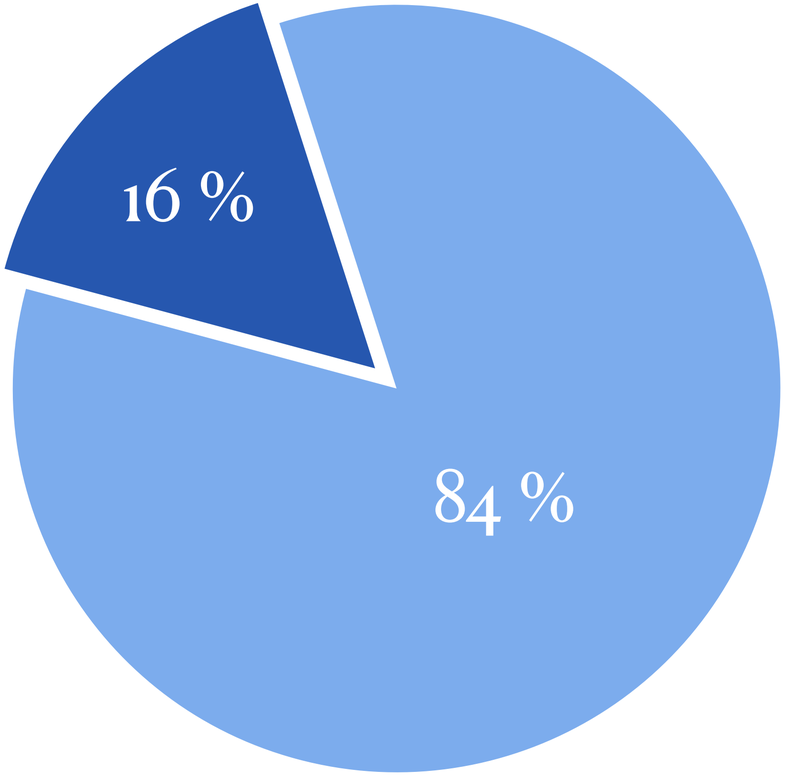
- Filter for „News“ in Title: 43 Shows
- Other filters like „Morning“, „Daily“ or „Report“ did not lead anywhere
- Total usable Shows: 49
- From 96 Stations

Shows


1100
990
880
770
660
550
440
330
220
110
0
Total Shows: 1010
„News“ in Title: 43
Usable Shows: 49

So from **91.310.616** identifiable snippets ...

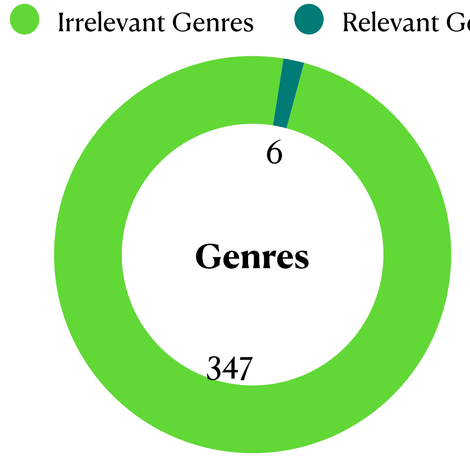
- **14.496.719** are usable in a news-context
- While **76.813.897** are not because they are too spontaneous or improvised



Spotify Podcast Dataset

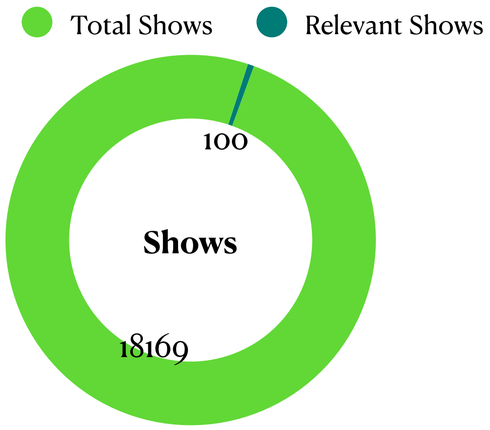
* 100.000 podcast episodes with aligned ASP transcripts
* more than 47.000 hours of transcribed audio
* automatic transcripts
* Word Error Rate: 18.1%
* has punctuation

Relevant Data

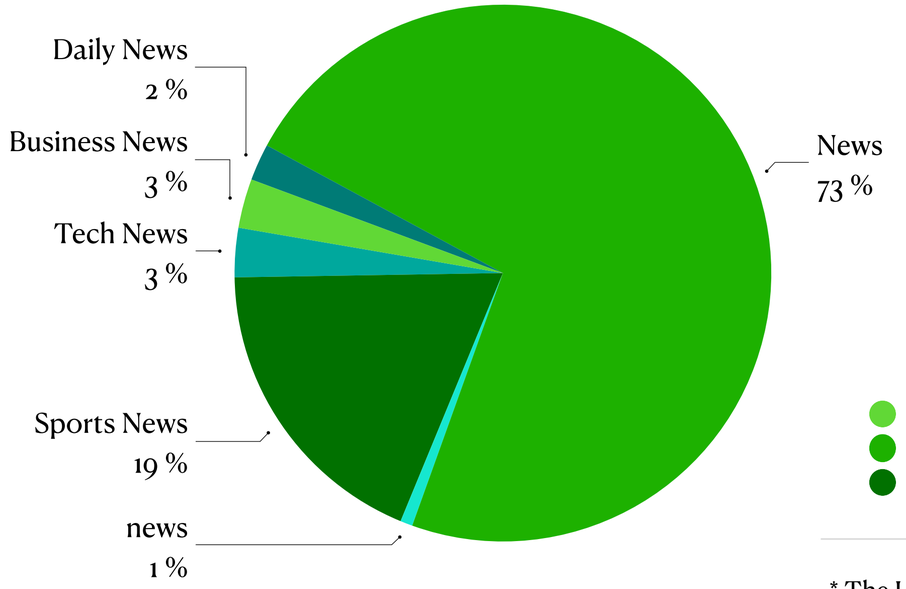


Relevant Genres

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

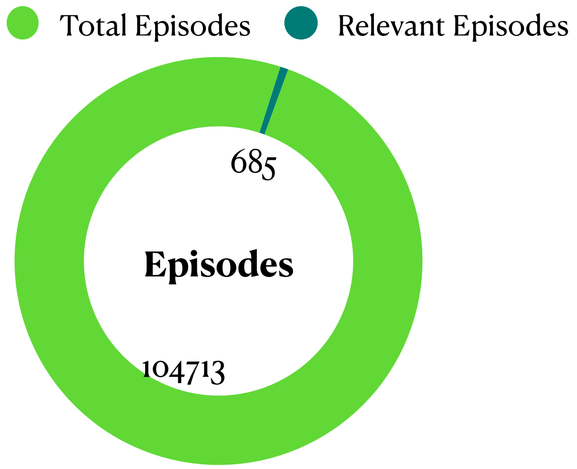


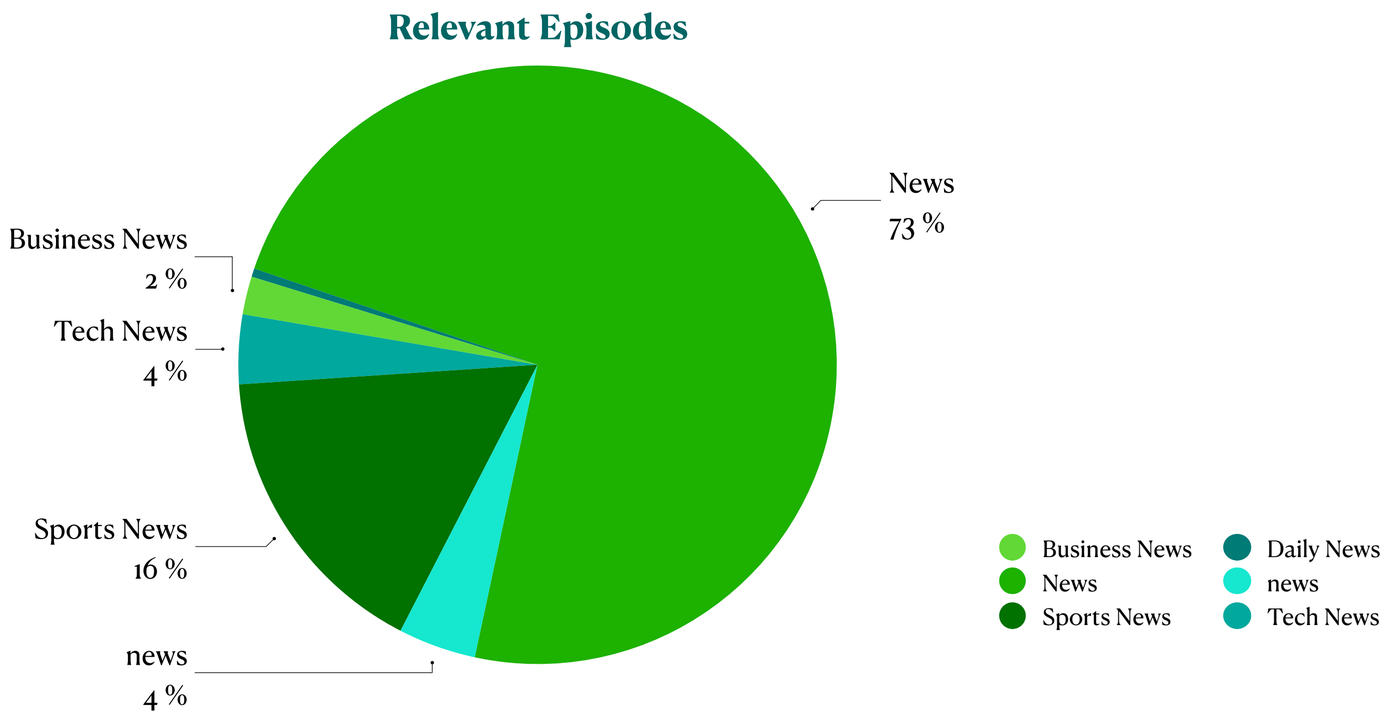
Relevant Shows



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

Episodes





general episode data

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

Quelle: The Spotify Dataset Paper

* 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

Index of Orality - Idea

Documentaries, Science- and History Shows are scripted and (highly) researched shows that are written for audio. Maybe these types of shows could provide an additional database?

They also have the same Index of Orality we want to achieve (see page 8):

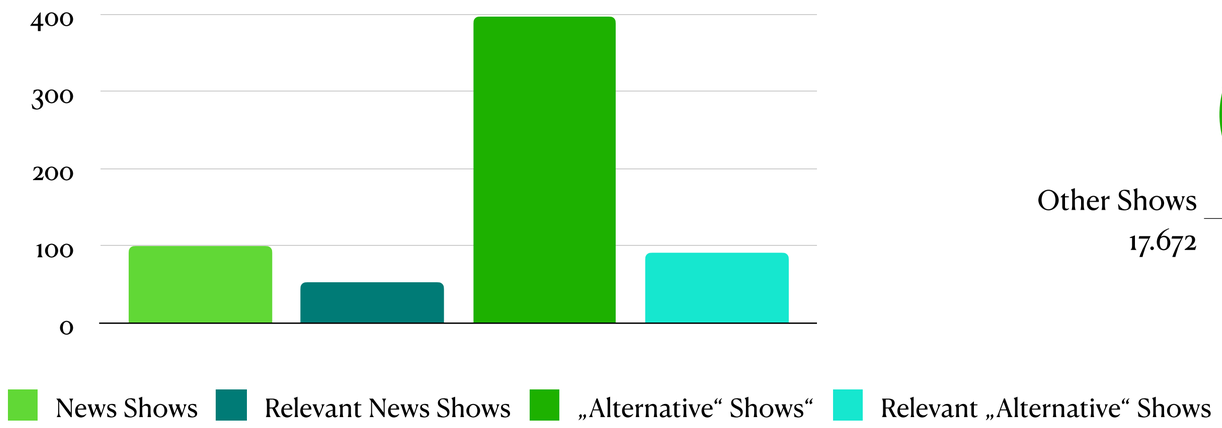
- many Participants = -1
- monologue = -1
- asynchronous production = -1
- synchronous reception = 1

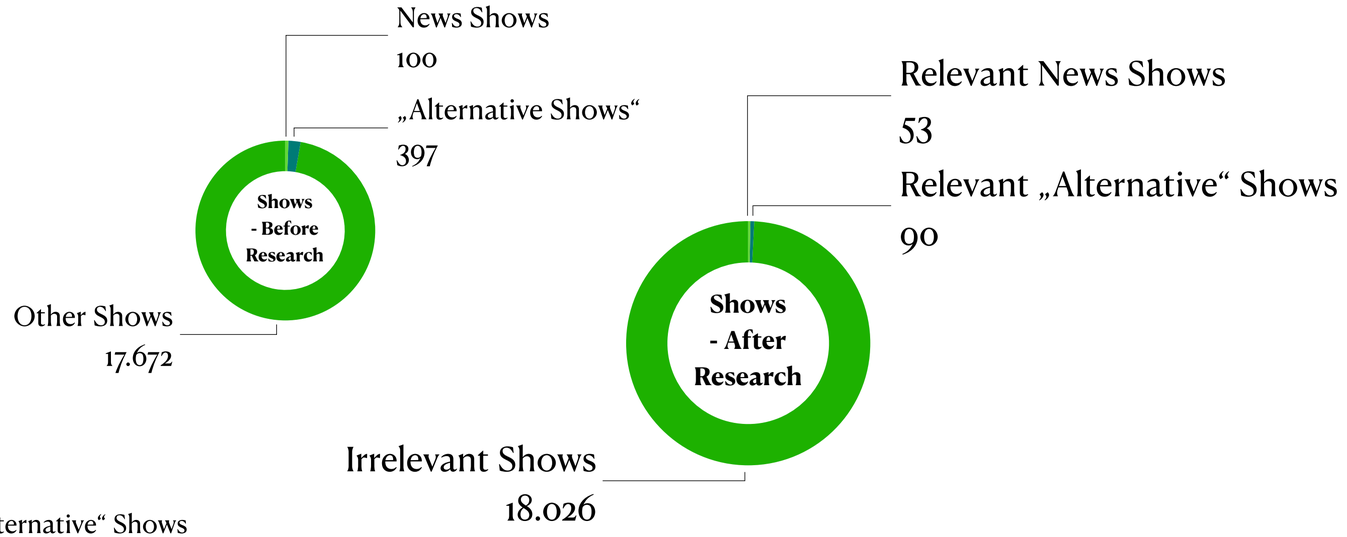
Index of Orality = -2

Update

Including Genres Science, History, Documentary and True Crime

- 397 additional potential Shows
- after research: Total of 143 Shows (incl. News)





Relevant Data - After Research

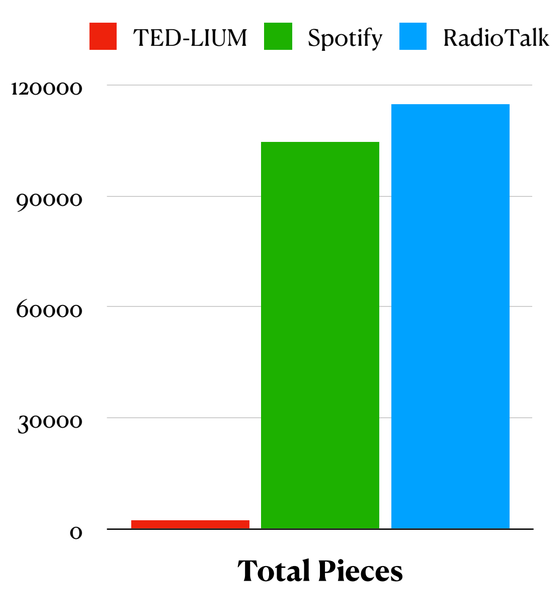
* Total of 143 Shows
* 2891 Episodes
* approx. 15.930.096 Words

TED-LIUM 3

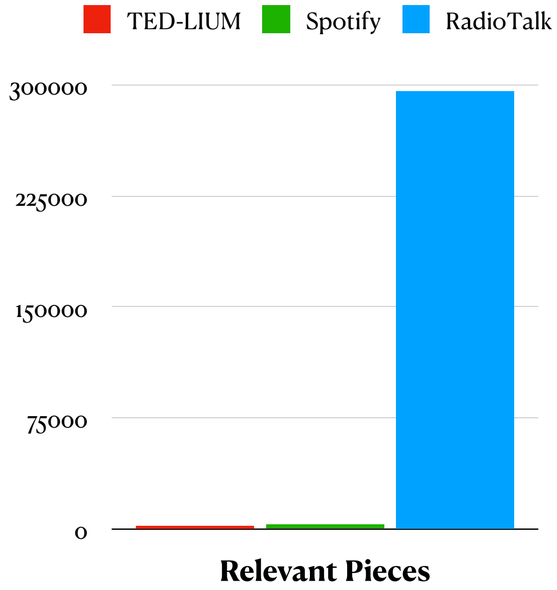
- 452 hours of transcribed speech
- 2351 speeches
- 4.9 M words
- automatic transcripts
- Word Error Rate: 6.7%
- no punctuation
- no genre information

Comparison

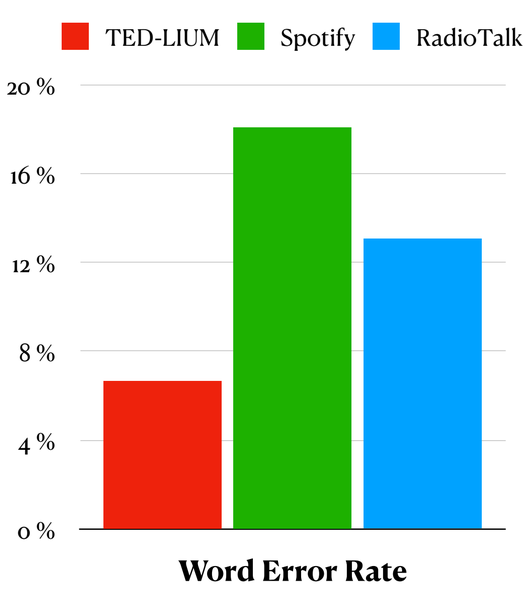
Comparison



Total Pieces



Relevant Pieces



Word Error Rate

Checklist

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