RadioTalk Corpus

- Total Entries: 115.769.559
- Total Shows: 1010
- Total Callsigns: 184

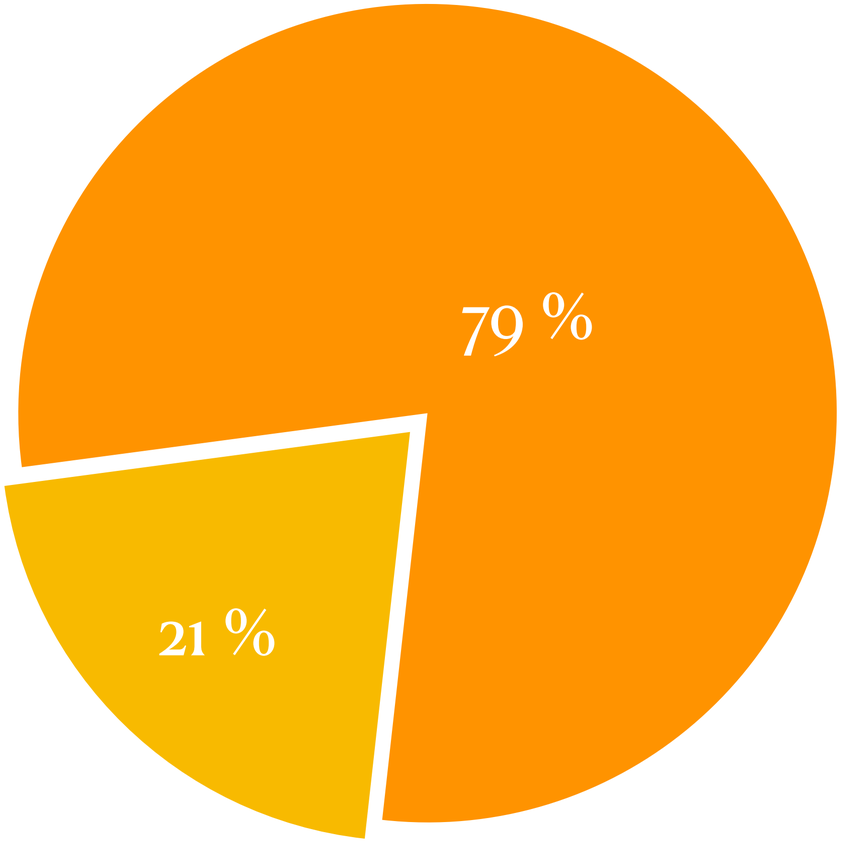
```
{
"content":"The transcribed speech from the snippet",
"callsign": "The call letters of the station the snippet aired on",
"city":"The city the station is based in, as in FCCC filings",
"state":"The state the station is based in, as in FCCC filings",
"show_name":"The name of the show containing this 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)"
}
```

Question: Which data can we use?

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

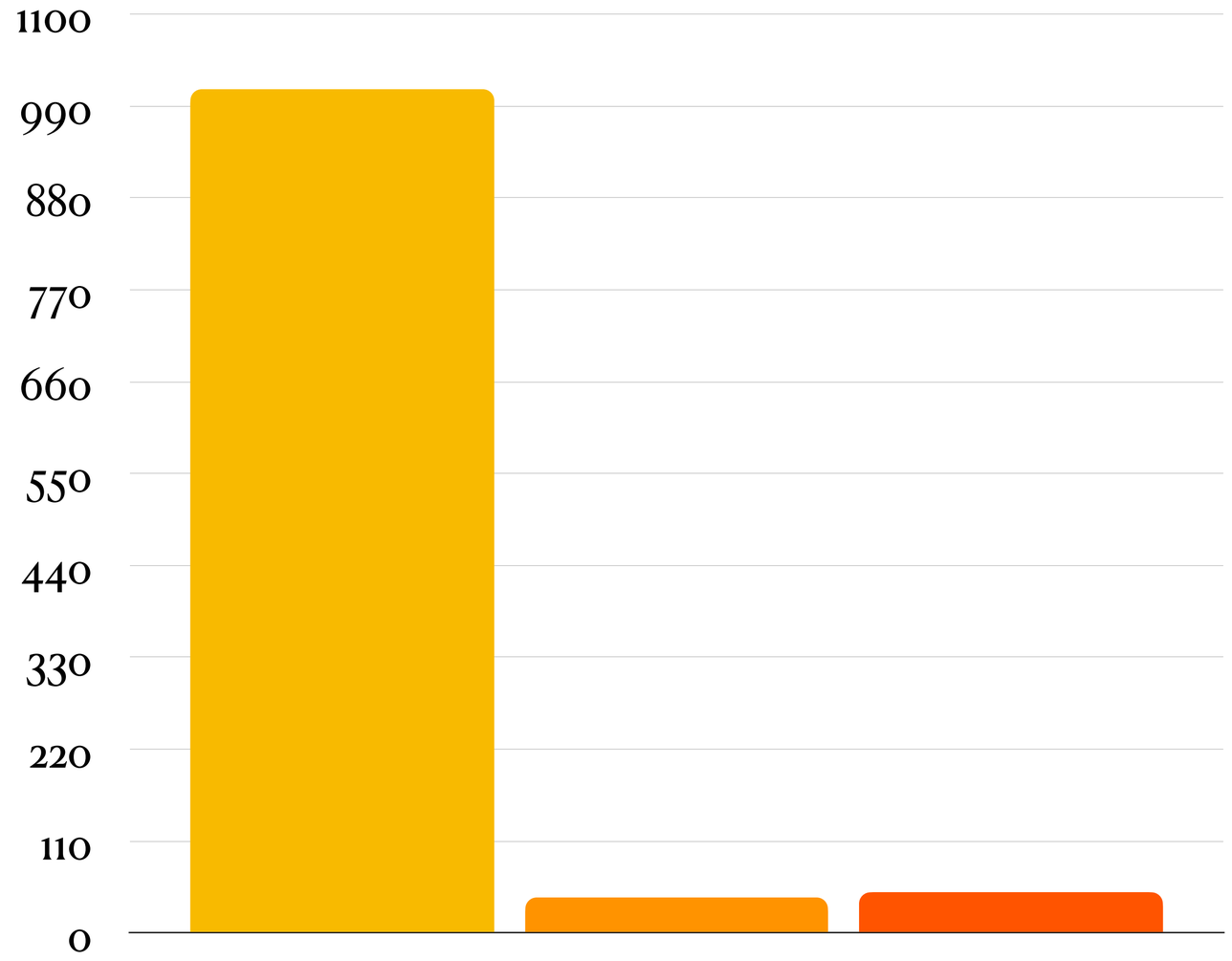
From a total of **115.769.559** snippets ...

- **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 really lead anywhere
- Total usable Shows: 49
- From 96 Stations

- Total Shows: 1010
- Shows with „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