Mining the Spoken Wikipedia for Speech Data and Beyond

Universität Hamburg • Department of Informatics
Natural Language Systems group
Arne Köhn • Florian Stegen • Timo Baumann

The Spoken Wikipedia project unites volunteer readers of Wikipedia articles. Hundreds of spoken articles in multiple languages are available to users who are – for one reason or another – unable or unwilling to consume the written version of the article.

We turn this speech resource into a time-aligned corpus, making it accessible for research and to foster new ways of interacting with the material.

Automatic Alignment Pipeline

Wikipedia Downloader

scrapes the spoken article category, checks for not yet downloaded articles and based on the template found in each such article:

- downloads audio, WikiText and HTML of the version read
- stores the article's meta-data found in the template and other sources

The Downloader can be **configured to match** templates and customs in **different languages** (so far: German, English, Dutch)

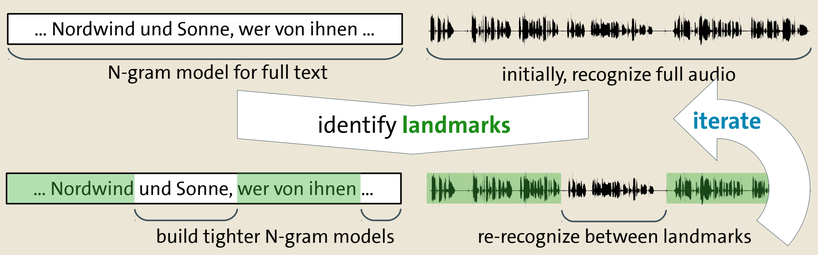
Text Extraction and Normalization

WikiText is extremely hard to parse. Therefore we use MediaWiki's HTML output and strip tables, boxes and other stuff[[citation needed]] that is not (or unpredictably) read out in the spoken version.

Our tool **uses MaryTTS** [1] for sentence **segmentation** and **tokenization**, and for **pronunciation normalization** (adding some rules, in particular for formulae and years). We also add the textual header that is read out before the actual content of the article.

Audio Alignment

We use a variation of SailAlign [2] as implemented in Sphinx-4 [3] with some modifications. SailAlign treats **alignment as repeated and successively more restricted speech recognition** using N-gram models derived from the text that is expected to be spoken. Correct stretches are used as landmarks to restrict re-recognition to tighter models.



SailAlign is **very robust to text/audio mismatches** which is a strong requirement given the modest data quality:

- text-speech mismatches (including text not read, audio not in text)
- mis-matching revision IDs (even leading to wrong articles)
- mis-normalization („*“ → „geboren“ or „Sternchen“?)

Favor quality over quantity in the alignments: rather leave out timings for dubious cases (e.g. mis-normalized tokens) than providing a full alignment that is partially faulty.

Bootstrapping Acoustic Models

Alignment requires acoustic and pronunciation models.
→ Sphinx 5.2 PTM acoustic models with good results for English.
For German, we used limited-quality acoustic models [4] and **iteratively built new models trained on the aligned data** found so far.
→ align more and more data as model quality improves.

We cross-checked the quality of the resulting models on the Kiel Corpus of Read Speech. We found that **alignment quality continues to increase**. We expect this effect to carry over to ASR.

Download our Software!

Our modular software consists of C# and Java code bound together with Python and Shell and works on (at least) Linux and Windows.
See QR-Code or paper for the link.

References

[1]: Schröder, M. and Trouvain, J. (2003): „The German Text-to-speech synthesis system MARY,“ *International Journal of Speech Technology*, **6**(3):365-377.
[2]: Katsamanis et al. (2011): „Sailalign: Robust long speech-text alignment,“ *Procs. of the Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.
[3]: Walker et al. (2004): „Sphinx-4: A flexible open source framework for speech recognition,“ Technical Report, Sun Microsystems.
[4]: Baumann et al. (2010): „InproTK in Action: Open-Source Software for Building German-Speaking Incremental Spoken Dialogue Systems,“ *Proceedings of ESSV*.

Spoken Wikipedia

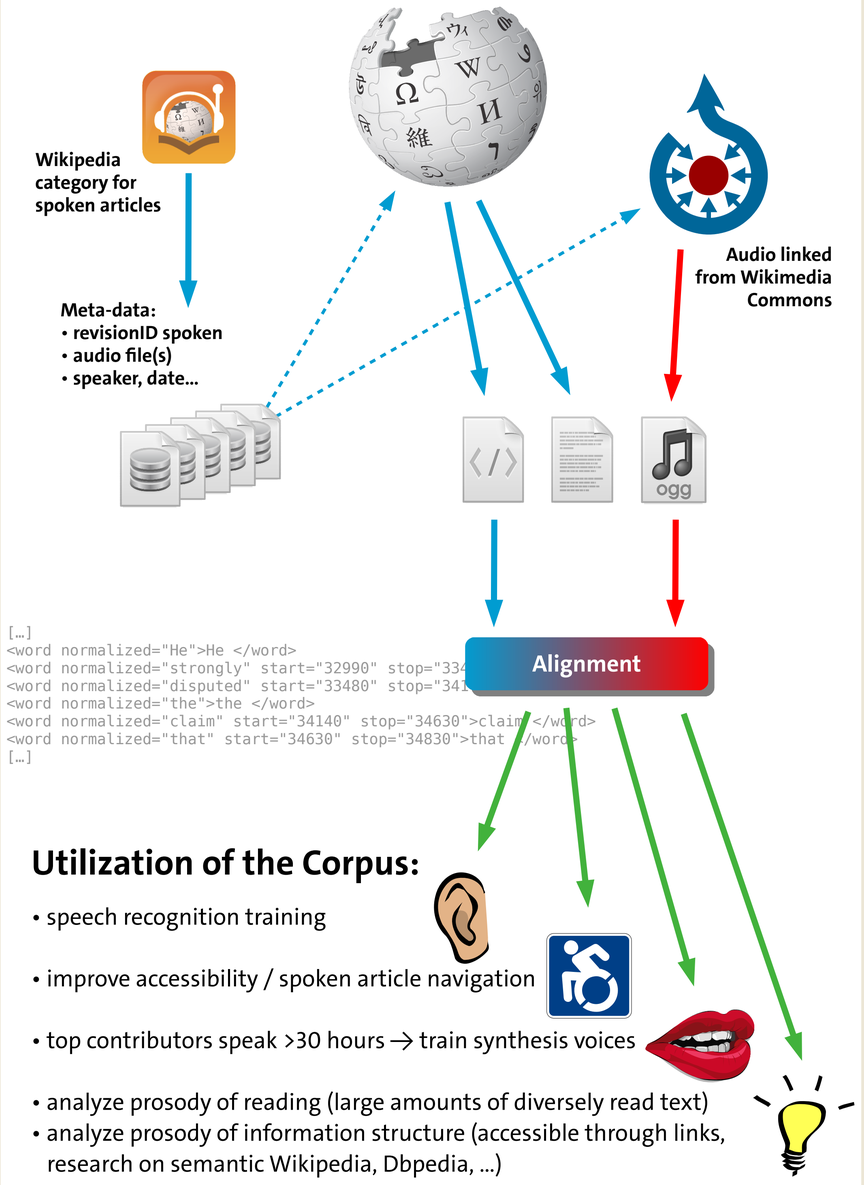
The Spoken Wikipedia is active in at least 28 language communities, with English, Dutch, German and French being the largest collections.

Articles are spoken by large and diverse sets of people, cover a variety of topics, focus on high-quality articles, and use a permissive license (CC-by-SA).

Category and template for spoken articles

All spoken articles use a template with slots: filename, speaker, date and revision spoken, ... to insert the audio player and a display of the meta-data on the Wikipedia page. The template also adds spoken articles to a root category.

Templates, categories, and meta-data vary between language communities!



Utilization of the Corpus:

- speech recognition training
- improve accessibility / spoken article navigation
- top contributors speak >30 hours → train synthesis voices
- analyze prosody of reading (large amounts of diversely read text)
- analyze prosody of information structure (accessible through links, research on semantic Wikipedia, Dbpedia, ...)

Download the Corpus!

For each article, the corpus contains:

- audio file(s)
- original WikiText
- HTML generated by MediaWiki
- cleaned and normalized text
- alignment between text and audio
- meta-information (who, when, what)

Languages: German and English
We are actively working on more data and more languages.

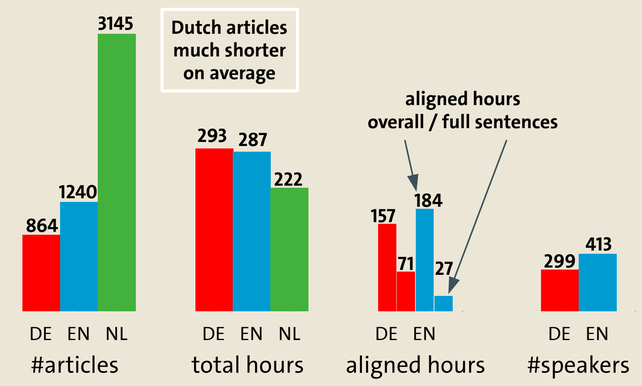
→ find it at:

http://nats-www.informatik.uni-hamburg.de/SWC/

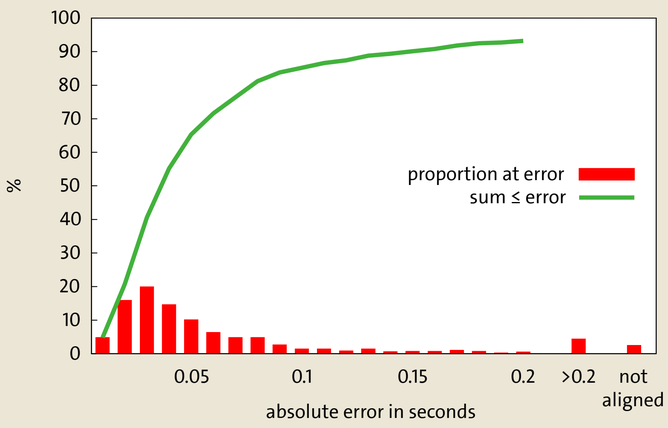
Our work, our corpus and this poster is CC-by-SA. Thanks to the Wikipedia contributors, as well as to Daimler-Benz-Foundation for funding.

Descriptive Metrics

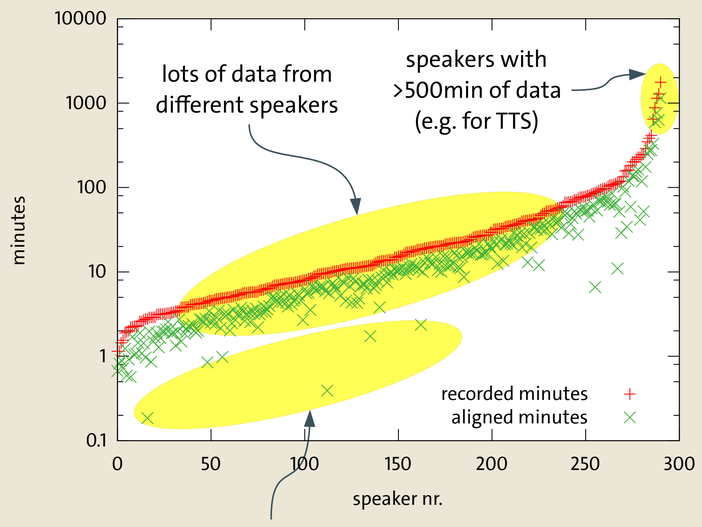
Overall statistics on the Spoken Wikipedia



Error to hand-corrected alignment (DE)



Recorded and aligned minutes per speaker (DE)



Low alignmend rates already mostly fixed (wait for next release)

Contributors speak related/interlinked articles: