LRE Journal manuscript No.
(will be inserted by the editor)

The Spoken Wikipedia Corpus Collection

Harvesting, Alignment and an Application to Hyperlistening

Timo Baumann · Arne Köhn · Felix Hennig

Received: date / Accepted: date

Abstract Spoken corpora are important for speech research, but are expensive to create and do not necessarily reflect (read or spontaneous) speech 'in the wild'. We report on our conversion of the preexisting and freely available Spoken Wikipedia into a speech resource. The Spoken Wikipedia project unites volunteer readers of Wikipedia articles. There are initiatives to create and sustain Spoken Wikipedia versions in many languages and hence the available data grows over time. Thousands of spoken articles are available to users who prefer a spoken over the written version. We turn these semi-structured collections into structured and time-aligned corpora, keeping the exact correspondence with the original hypertext as well as all available metadata. Thus, we make the Spoken Wikipedia accessible for sustainable research. We present our open-source software pipeline that downloads, extracts, normalizes and text-speech aligns the Spoken Wikipedia. Additional language versions can be exploited by adapting configuration files or extending the software if necessary for language peculiarities. We also present and analyze the resulting corpora for German, English, and Dutch, which presently total 1005 h and grow at an estimated 87 h per year. The corpora, together with our software, are available via `http://islrn.org/resources/684-927-624-257-3/`. As a prototype usage of the time-aligned corpus, we describe an experiment about the preferred modalities for interacting with information-rich read-out hypertext. We find alignments to help improve user experience and factual information access by enabling targeted interaction.

Keywords Wikipedia · speech corpus · found data · annotation · robust text-speech alignment · spoken hypertext · eyes-free speech access

This article extends and consolidates previous research (Köhn et al, 2016; Rohde and Baumann, 2016). This work was partly supported by a PostDoc grant by Daimler-and-Benz-foundation to the first author.

T. Baumann
Carnegie Mellon University, School of Computer Science, Language Technoloy Institute, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. E-mail: tbaumann@cs.cmu.edu

A. Köhn · F. Hennig
Universität Hamburg, FB Informatik, Natural Language Systems group, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany. E-mail: {koehn,3hennig}@informatik.uni-hamburg.de

1 Introduction

Most time-aligned speech data for corpus analyses or estimating model parameters (e. g. for speech recognition of synthesis) is non-free, producing a barrier for research. In addition, models generated from these corpora cannot be freely distributed which further limits research. Free speech corpora such as *Voxforge*[1] only contain limited amounts of data – e. g. 56 hours for German, 130 hours for English and 10.5 hours for Dutch at the timing of writing – and crucially, these resources only grow with input from entities interested in speech processing but do not get input from the general public.

The *Spoken Wikipedia*[2], in contrast, is a branch of the Wikipedia project and serves the main purpose of enabling the aural consumption of Wikipedia contents for persons who are unable or unwilling to read (out of alexia, visual impairment, or because their sight is currently occupied, e. g. while driving). It is constantly expanding and evolving and is of considerable size for several languages. It is built by volunteers, presenting a truthful picture of what self-motivated individuals 'in the wild' consider useful read speech. Only as a side effect, the Spoken Wikipedia is a large collection of factual read speech covering a broad variety of topics under a free license with corresponding text available. Our research goal is to turn this data into a structured corpus, and to build sustainable software that allows to update and extend the corpus in the future as the underlying Spoken Wikipedia grows and evolves.

Wikipedia is accepted as the standard source for encyclopedic knowledge on the web and is ranked as one of the 10 most heavily accessed websites on the Internet[3], a further indicator towards the relevance of Wikipedia content. The (written) Wikipedia has already been widely used for research in almost all areas of computational linguistics, including but not limited to lexicology (Grefenstette, 2016), lexical simplification (Horn et al, 2014), named entity recognition (Nothman et al, 2009), named entity relation mining (Strube and Ponzetto, 2006; Iftene and Balahur-Dobrescu, 2008), co-reference (Ghaddar and Langlais, 2016), question answering (Ahn et al, 2004; Buscaldi and Rosso, 2006), general language understanding (Hewlett et al, 2016), building taxonomies (Laura Kassner and Strube, 2008), or parallel corpora for machine translation (Tufiş et al, 2014). Beyond the articles, meta data has also been used, including article category graphs (Zesch and Gurevych, 2007), article histories (Max and Wisniewski, 2010; Wijaya et al, 2015), and discussion threads (Prabhakaran and Rambow, 2016). Other work aims directly at improving the Wikipedia, for example discovering missing inter-language links (Lefever et al, 2012), classifying edits (Yang et al, 2016), or detecting vandalism (Potthast et al, 2008). Finally, the Wikipedia project is of course also a subject in the social sciences (Stegbauer, 2009; Suh et al, 2009, and many others).

In contrast, the Spoken Wikipedia has previously hardly been used in research, although it is a broad and multi-lingual source with lots of automatic and manual annotation available (such as links and topic relatedness) for the textual material. Fur-

[1] `http://www.voxforge.org`

[2] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia`; also contains links to other languages. Please note that the authors are not involved in that project.

[3] `http://www.alexa.com/topsites`

thermore, the tools and annotations available for the written Wikipedia (as presented above) are directly applicable to the textual side of the Spoken Wikipedia, making the genre and actual texts of the Spoken Wikipedia particularly well-researched. The major problems for leveraging this wealth of information is the missing linkage between the speech and text as well as the semi-structured and noisy nature of Wikipedia data. We attack this issue with a novel annotation schema that retains an exact correspondence of the annotated data with the original hypertext.

We present an open-source software pipeline that downloads and extracts the data, and aligns speech in the Spoken Wikipedia to the article text being read. The tool works for multiple languages and can easily be extended to cover more. We show how we are able to bootstrap free speech recognition models and how these bootstrapped models increase the coverage of time-alignments in our corpus for Dutch.

We test the utility of a time-aligned hypertext corpus by analyzing the usability of spoken hypertext as an exemplary research question. We build a preliminary (graphical and voice-based) interface to the Spoken Wikipedia that uses time alignments for improved search and navigation. We find that users are efficiently supported by the time alignments and that users find information much more quickly than when listening to the non-interactive Spoken Wikipedia alone.

After a description of the structure of the Spoken Wikipedia and some statistics in Section 2, we present our main contributions:

- We present an annotation schema as well as open-source software for bringing Spoken Wikipedia collections into a structured and time-aligned form while retaining an exact correspondence with the source hypertext. (Section 3)
- We present and analyse, for three languages, the resulting corpora of read encyclopedic content read by a large variety of speakers, available under an free license, and roughly a magnitude larger than previously existing freely licensed spoken collections of factual texts. (Section 4)
- We investigate navigation behaviour in read-out hypertext – which we call *hyperlistening* – and present the finding that hyperlistening is greatly beneficial for factual information access as compared to listening to a standard recording. (Section 5)

We conclude in Section 6 in which we also point to related work that already uses the Spoken Wikipedia Corpora as well as outline further research opportunities.

2 The Spoken Wikipedia

The Spoken Wikipedia is a project in which volunteers read out and record Wikipedia articles (which are originally produced in writing). According to the project's mission statement,[4] the main aim is to provide high quality aural access to Wikipedia, e. g. for persons with visual impairments, and for other hands- and eyes-free uses, such as while driving. All audio data is available under a Creative Commons Share-alike license, the same license the Wikipedia uses for article texts.

Spoken data is available for 30 languages, with English, German and Dutch by far being the largest collections. We focus on German in this article as there is otherwise

[4] `https://en.wikipedia.org/wiki/File:Spoken_Wikipedia_Benefits.jpg`

relatively little free speech material available for German, but we also validate our generalizations by looking at English data. Our software pipeline described below also works for Dutch (with some limitations) and can be extended for other languages or Wikipedia varieties (such as simplified English).

2.1 The Organization of a Wikipedia Article

All Wikipedia articles follow the same overall structure:[5] the article starts with a short summary section followed by the main content which is divided into sections and subsections. The main content is followed by appendices containing lists of related articles, endnotes and references, pointers for further reading and finally external links. This is followed by categorization information; some content may be presented in 'info boxes' (e. g. information that similarly appears in multiple articles, such as for all countries: a flag, an anthem, the GDP and number of inhabitants). The body of an article may also contain tables, figures and images and their respective captions.

Wikipedia articles are rendered into HTML from *WikiText*, which is a simple-to-write markup language for hypertext, enabling *inter alia* sectioning, links (within and beyond the Wikipedia), endnotes (used for citation), categorization, templating and the integration of multimedia files. As per the style guidelines, hyperlinks in the main part of the article must link to other Wikipedia articles and each related lemma should only be manually linked on the first or most important occurrence in the text (links to other lemmata are not set automatically by the rendering system). Categories (some of which form a hierarchy) are simple markers that are added to the WikiText to group related articles. Info boxes are specified using a parameterized templating mechanism. Articles may call templates specifying the attribute/value pairs to be used by the template and the templates specify how the given parameters are to be displayed (e. g. amended with other template specific information), and potentially add calling articles into specific categories.

As a result of templating, the rendering of an article into HTML is not determined by the WikiText of the article alone, but also depends on the templates used. As the Wikipedia evolves over time, articles change. Old states can still be referenced by their `oldID` and through the article's history. However, there does not seem to be a way of referencing old versions of the included templates (which, for the most part, create info boxes that are often not read out). Thus, recovering the former rendering of an old revision of an article is non-trivial. As articles are presumably most often read from an on-screen browser presentation, there remains some uncertainty as to what exactly has been seen by the reader when she read the article. All of this calls for a robust text-speech alignment procedure (see Section 3.4).

2.2 The Organization of the Spoken Wikipedia

The organization of the Spoken Wikipedia predetermines our extraction process (as described in Section 3.2) and is reflected in Figure 1.

[5] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Layout`

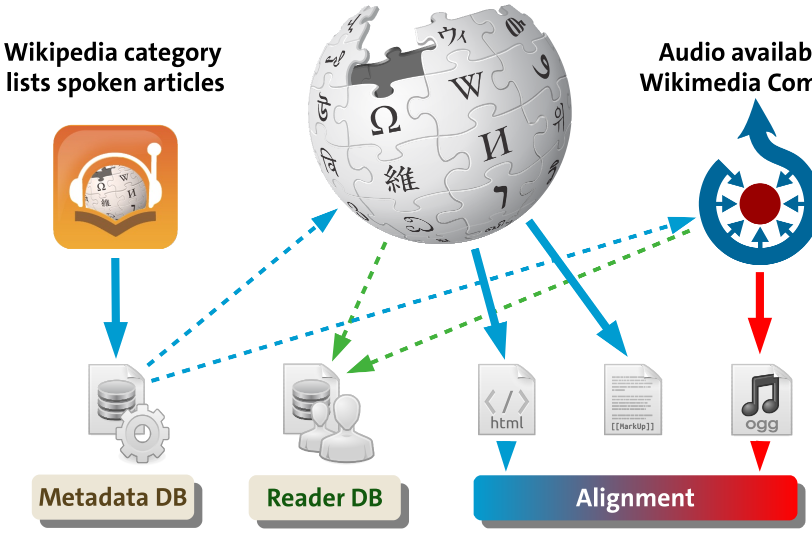

Wikipedia category lists spoken articles
Audio available at Wikimedia Commons
html
[[MarkUp]]
ogg
Metadata DB
Reader DB
Alignment

Fig. 1: Depiction of the structure of the Spoken Wikipedia and the corresponding process flow during corpus creation.

Every language version is organized as follows: First a volunteer reads, records and verifies an article and this step sometimes involves some textual editing as a result of careful reading by the volunteer. The volunteer then uploads the audio file to Wikimedia Commons and includes speaker meta information (e. g. about gender, dialect, or recording equipment) in a free-text format. For long articles, the recording may be split into several files which are all uploaded and cross-referenced on Commons.

The volunteer finally adds a template to the WikiText of the article that contains multiple parameters such as: the link to the audio file(s), the date and/or `oldID` when the article was read, and additional speaker properties (language dependent). The format of the template and the parameterization is language-specific. The template in the WikiText is at runtime expanded by the MediaWiki engine into a particular type of info box which displays a widget for playing the audio, links to the article revision read (if available), and reports the date read and the reader.

Finally, the template also adds a category marker to the article that marks the article as being part of the Spoken Wikipedia article collection in the given language. All spoken articles in a language can then be found by querying for this category. Most languages also contain ‘portals’ to their spoken content which may contain e. g. a curated categorization of spoken articles, or instructions on how to contribute.

2.3 The Articles of the Spoken Wikipedia

The Spoken Wikipedia has several advantages: articles are read by a large and diverse set of people, and are not only available free of charge but also under a free license

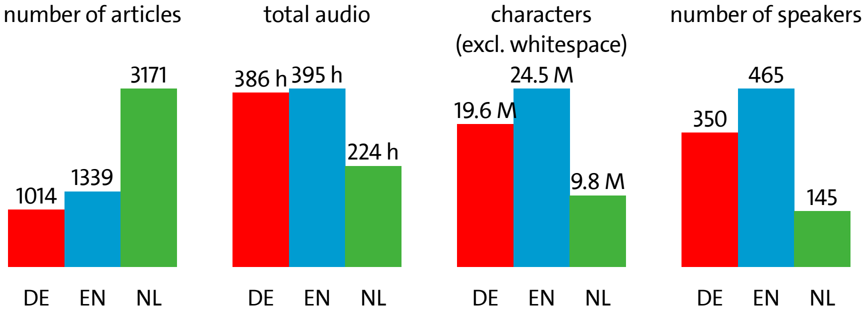


Fig. 2: Some key statistics about the Spoken Wikipedia in different languages.

(CC-by-SA). Although the Wikipedia is constantly evolving, the audio files are co-referenced with the exact article revision that was being read (with some exceptions, see below), allowing to match the audio with the read text. In this sub-section we look at some key statistics of the Spoken Wikipedia.[6]

Figure 2 reports the overall sizes of spoken data for three languages. The English Spoken Wikipedia is the largest collection (by audio duration) and so far contains 1339 pages read by 465 identified speakers (speaker information is missing for some speakers; we use fuzzy matching to combine same speaker names formatted in different ways), totaling 395 hours of audio.

The German Spoken Wikipedia contains 1014 spoken articles read by 350 identified speakers (speaker information is missing for some articles), totaling 386 hours of audio (of which 38 hours are missing speaker information). A plot of speaker contributions is presented in Figure 3 (upper red dots). The average duration read per speaker is 54 minutes but as that figure shows, the distribution of contribution (by audio duration) is highly skewed. Very few speakers speak tremendous amounts of audio (up to 60 hours).[7] The median reading time is 13 minutes and the 25 % and 75 % quantiles are at 8 and 40 minutes, respectively. Although many speakers only read a single article (not emphasized in Figure 3), most are still represented by a fair amount of audio.

As can be seen in Figure 2, the number of articles spoken in the Dutch Wikipedia by far exceeds the other languages, although in total there is less audio. This reflects differences in what has been read: while the average audio per article is 37 minutes for German and 29 minutes for English, the average recording is only 7 minutes per article in Dutch. In other words: while very complete articles have been read for English and German, Dutch spoken articles most often only provide a brief overview of a topic.

As a means of text-based comparison to other corpora (spoken or written), we report the number of non-whitespace characters contained in the Spoken Wikipedia

[6] We want to point out that Wikipedia itself does not offer such statistics (beyond the number of articles in the category) but that such statistics can be easily created with the software described in Section 3.

[7] We have not manually checked yet whether that speaker may have also edited other speaker's work and uploaded this under one unifying name.

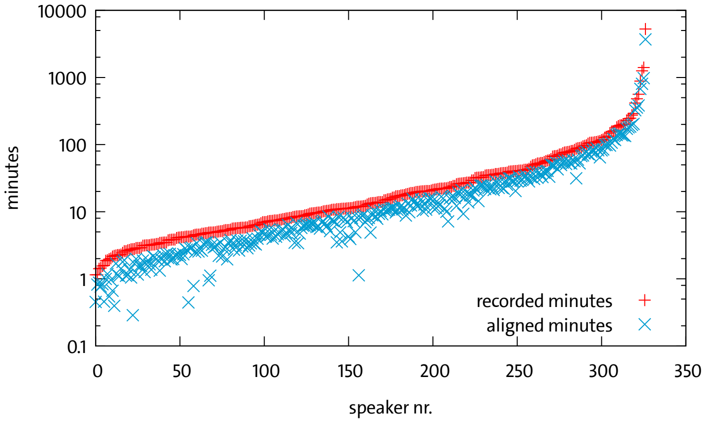


Fig. 3: Contribution of audio by speaker in the German Spoken Wikipedia. Half of the speakers contribute between 8 and 40 minutes of audio; only a few contribute (much) more (up to 60 hours); except for some outliers, the alignment rate (see Section 4) is relatively stable across speakers.

Table 1: Comparative statistics of spoken and written versions of the German and English Wikipedia (as of 2016-06-15).

		German	English
Written	# articles	1,950,022	5,174,458
	— distinguished	6,283	29,189
	average text size	5.3 kB	6.2 kB
Spoken	# articles	916	1,344
	— distinguished	314	213
	average text size	25.8 kB	26.0 kB
Spoken Coverage	articles	0.047 %	0.026 %
	— distinguished	5.0 %	0.73 %
	est. speech time	0.22 %	0.11 %

language versions in Figure 2. We believe this number to be more easily comparable across languages than the number of word tokens, at least for latin script-based languages (however, see Table 2 for per-language token counts).

Wikipedia contains millions of articles on all sorts of topics in the major languages, inviting the question of whether the Spoken Wikipedia's meager thousand articles per language (at least for English, German and Dutch) are of any practical relevance beyond being 'nice to have' speech data for speech and NLP research purposes.

To quantify this concern, we compare the composition of the written and spoken collections for German and English in Table 1. As can be seen in the table, both language versions consist of several million articles each, with a small proportion

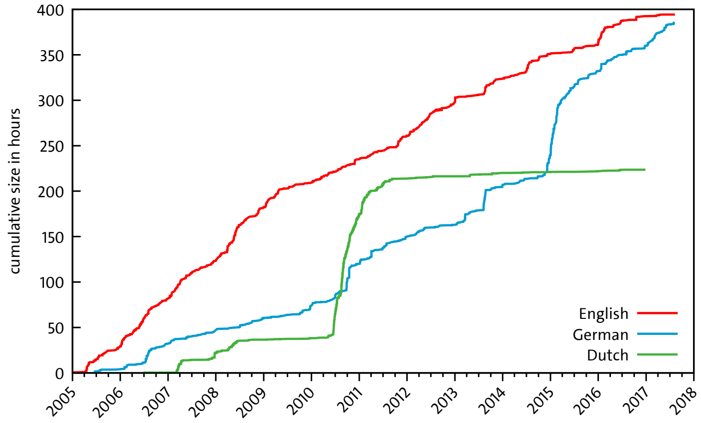


Fig. 4: Growth of the Spoken Wikipedia over time. Overall growth for the three languages combined is at 87 hours per year.

of *distinguished* articles.[8] We estimate the average length of written articles on a random sample of 1,000 articles for both languages (using –for simplicity– their size in bytes as a proxy for text length). We find that articles selected for being spoken are (a) much longer than average articles (4-5 times as long), and (b) more often come from one of the distinguished article categories. In the German Wikipedia, some 5 % of distinguished articles have been read. Nevertheless, only a tiny proportion of the full Wikipedia is available as a naturally read version (0.11-0.22 %) and we estimate that a fully read Wikipedia would have an audio duration of several decades – indicating that full coverage may be hard to achieve and to maintain (yet, the goals of the project outlined by Andersson et al (2016) is to considerably simplify the recording of articles).

Given the voluntary nature of Wikipedia, articles being read are not randomly selected but depend on the interests of the readers and their perceived importance of the article. They span a wide variety of topics such as (for German): cities (Ingolstadt, 152 minutes), famous researchers (Carl Friedrich Gauß, 54 minutes), technical articles (Microsoft Windows NT 3.1, 67 minutes), or sciences (number theory, 22 minutes).

Figure 4 shows the growth of the Spoken Wikipedia over time (and per language). As can be seen, there is sustained growth of the projects over more than 10 years, at least for German and English, and growth could be described as linear with a growth rate of about 5-5.5 minutes of audio per day (or 32-33 hours per year).[9] The

[8] English articles can be distinguished as either ‘good’ or ‘featured’, where the corresponding German categories are ‘lesenswert’ (worth reading) and ‘exzellent’.

[9] The vertical steps in the graph, particularly for English, appear to be a measurement artifact: audio files were moved in bulk and received newer timestamps. The spurt for Dutch, however, seems to have been a real, concerted effort.

linear growth pattern is in line with the overall growth (in total bytes) of the written Wikipedia.[10]

As the Wikipedia evolves, spoken articles may become outdated and volunteers may want to re-record a new version of an already spoken article. Although this practice is endorsed in the guidelines, we could find only twelve German articles that have been re-read (as the community still focuses on adding articles rather than adapting existing ones). Half of the articles have been re-read by other speakers, the other half has been re-read by the original speaker. On average, 3.7 years expired before an article was re-read.

3 Corpus Preparation Pipeline

Our processing pipeline software works as follows: First, we scrape the Wikipedia for articles that are in the category for spoken articles. We then download the corresponding text and audio. The text is then normalized before it is aligned to the audio.

One of the main challenges is the constant evolution of Wikipedia: articles (or their spoken versions) are added, article text is revised, meta-information is updated or erroneously breaks for a multitude of reasons. As a consequence, our solution must not rely on manual corrections of the extracted data. Instead, our solution needed to be robust to errors, able to extend the corpus without forcing recomputations on the unchanged data, and provide systematic workarounds for missing data (until that data is added at the source).

We do not perform manual corrections on the generated data as the large amount of data as well as the ever-growing source would make this very expensive and hard to maintain. We instead rely on the algorithm being conservative with the alignment and evaluation on manually aligned data shows that it only misaligns very few words. In some cases, we have fixed problems and inconsistencies at the source, the article or audio meta-information pages and re-ran our experimental pipeline afterwards.

In this section, we describe the steps outlined above in detail, after first describing our annotation schema for the resulting data sets.

3.1 Annotation Schema

The Spoken Wikipedia Corpora are multi-modal corpora consisting of audio and text. The text is in an XML-compatible HTML5 format as downloaded via the Wikipedia API, which already contains textual structure such as sections and paragraphs and hypertext markup such as links that are worth keeping. The SWC annotations, however, align the text on a much more fine-grained level than given by the existing HTML elements – as fined-grained as individual speech sounds in the audio track of the corpus. In addition, sentence boundaries and tokens should be marked

[10] Unlike the total number of articles, which appears to saturate, the raw data contained in Wikipedia still grows linearly: `https://commons.wikimedia.org/w/index.php?title=File:Wikipedia_article_size_in_gigabytes.png&oldid=243766150`.

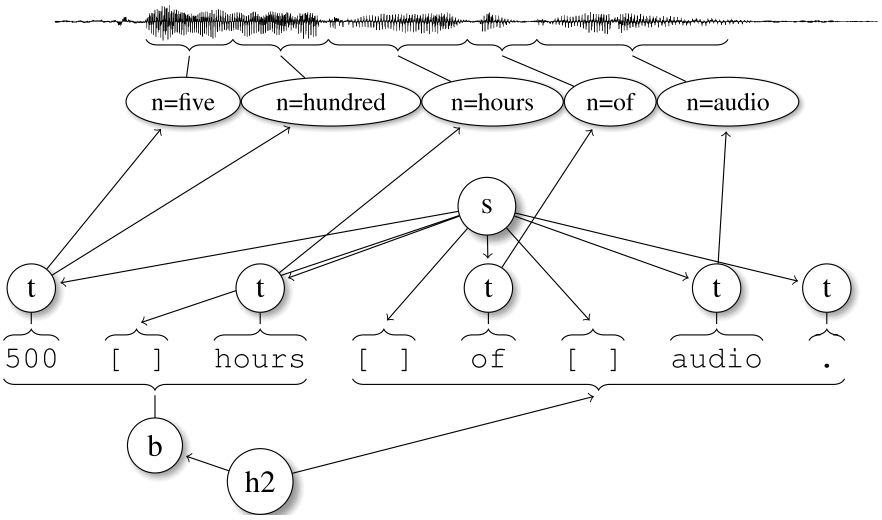


Fig. 5: Exemplary annotation of "**500 hours** of audio." with SWC annotation that binds text to audio above and HTML markup that adds hypertextuality below. The SWC annotation marks sentences (s), tokens (t) and adds normalization information (n), which refers to the audio. Note that the whitespace ([ ]) between words are original characters that are attached to the sentence but are not part of any token.

as produced by NLP tools. As a result, the original and SWC tree structures may conflict. There is a fundamental problem with nesting tags inline when conflicting tree structures are involved. For example, HTML links can span multiple word tokens while simultaneously spanning only parts of one (or more) of the tokens. As a concrete example, take `<t>Warschau-<a>Berliner</t> <t>Urstromtal</t></a>` in which tokenization and linking interfere.[11] The existing tree structure of the HTML cannot be integrated with a conflicting tree structure resulting from sentence segmentation and word tokenization and hence we opted against enriching inline the HTML format with additional annotation.

Using a stand-off annotation circumvents the problems introduced by a direct inline annotation and this approach is used e. g. by PAULA (Chiarcos et al, 2008). Stand-off annotations reference the original using XPointer expressions and/or XLink references. We evaluated XML-based stand-off annotations and noted that most XML libraries lack support for referring to text ranges in character data. This would however be needed due to the above-mentioned inability to provide fine-grained segmentation given the HTML tags. We opted against stand-off annotations as these are tied to a limited set of tools and we think that the SWC will be mostly used programmatically and will profit from a broad tool support.

[11] One might argue about the correctness of tokenization or appropriateness of linking. However, such theoretical considerations do not solve the problems associated with using preexisting data and tools.

To solve the problems outlined above, we interpret the original HTML as actual tree-form annotations of the underlying text characters (all material that is in between angular brackets, i. e. not the markup). This way, the SWC alignment and HTML markup can be treated as two different annotations for the same original text and they can be either consulted independently or merged via their correspondence. Merging the two annotations is only possible because the texts are *exactly* the same character-for-character (including all whitespace and line breaks – but not necessarily using the identical encoding scheme for Unicode characters in XML). Obtaining this exact match between the two annotations is not possible in preexisting annotation schemes as they only partially encode and preserve original whitespace (and sometimes also tamper with other material).

An example of our annotation scheme can be seen in Figure 5. In the example, the text `500` is identified as one token which is normalized to two items (`five` and `hundred`) which are individually aligned to the audio. As can be seen, some text is bold-faced by HTML (e. g. to indicate importance) and this does not clash with the SWC alignment. The full annotation format deals with additional material that is read (but not written) such as a license statement that is always spoken in the beginning of the audio, as well as to mark material that is written but usually not read (such as tables or lists of references), with titles, section headings and other specifics, and allows for segmental phonetic annotation. The specification of SWC XML is available as part of our software suite.

3.2 Scraping the Wikipedia

All articles with a spoken version contain a template which in turn results in some info box and a category marker. As a result, all articles with a spoken version are contained in the same category (named ‘Spoken Articles’ in the English version). We use the Wikipedia API to query the spoken article category, then examine the article source for the template, which contains the read article revision ID, speaker name and reading date, as well as a link to the Wikimedia Commons page that contains the actual audio (in one or more files, almost always encoded as OGG-Vorbis in high quality).

The Wikipedia is a semi-structured database, which means that values can be missing or mal-formatted. Our software contains many workarounds to deal with such issues (e. g. if the revision ID is missing from the template, we estimate it from the reading date and the article history available via the API).

Meta data about the recording conditions (e. g. microphone used or intermediate audio encodings), and the speaker (e. g. dialect, gender) can often be inferred from the template in the article, the page on Wikimedia Commons, or even by following a link to the user page of the speaker (which is often linked to from Commons). However, there appear to be few encoding conventions for such data or conventions are not being followed. We do not try (at present) to interpret all these data but we download and store all possible meta-data entries with the downloaded articles for posterior use.

3.3 Structure Conversion and Text Normalization

Wikipedia pages can be downloaded in several formats including WikiText, an HTML-like format (that is probably fed to the CMS), and an XML-representation that is somewhere inbetween. While we download all these formats, the HTML is easiest to convert to raw text, including stripping footnotes, "citation-needed" marks, and other Wikipedia markup and hence we base our annotation on this representation.

We use MaryTTS (Schröder and Trouvain, 2003) for sentence segmentation, tokenization and normalization. Our intermediate processing ensures that the original text and the final normalized text remain in synchrony so that timing information for the original text can later be inferred based on the alignment of the normalized text.

As a special kind of normalization, we add the spoken "header" of each article that mentions the name of the article, the license, the date it was read, and possibly by whom, using a pattern filled from the meta-data. We also filter out any textual lists of references, as these are typically not read (and would be hard to normalize).

3.4 Text-Speech Alignment

Starting from an audio file and the corresponding article text, we create an alignment between both. The key challenge regarding the alignment: Although we have an audio file of the article being read as well as the article's source code, the written part does not always match what is being read. There are (1) parts in the text that are not being read, (2) parts we don't know how they are being read, and (3) there are parts in the audio that are not part of the text.

Regarding problem (1), we mark regions of the text that are unlikely to be read such as footnotes, tables and info boxes to be excluded by the alignment process. Even if they are being read, it is usually not predictable in which manner and ordering they appear in the audio. Note that these regions – and the exclusion markers – are still preserved in the output to preserve exact correspondence to the source text. This approach might err on the side of excluding too much from the alignment process, but it prohibits false alignments to text that is not spoken at all.

Regarding (2), there is some text where we don't know whether or how it will be spoken, e. g. the headings and mathematic formulas. We try to normalize some formulas, but the coverage is limited (see also Ferres and Sepúlveda (2011) for a solution to this problem). In the case of unsuccessful normalization, the alignment algorithm will simply not be able to find an alignment.

Regarding (3), typically, the audio starts (and often ends) with a disclaimer about the origin of the text and the license of the audio. This is not part of the article but is sufficiently similar for each audio file that we can generate the text and prepend it to the text to be aligned. These artificially generated passages are again marked in the annotation to distinguish them from the proper article text. Other parts will not be aligned (as the corresponding text is not available).

To perform the audio alignment, we employ a variant of the SailAlign algorithm (Katsamanis et al, 2011) implemented in Sphinx-4 (Walker et al, 2004) with some extensions as described below. SailAlign treats audio alignment as repeated and

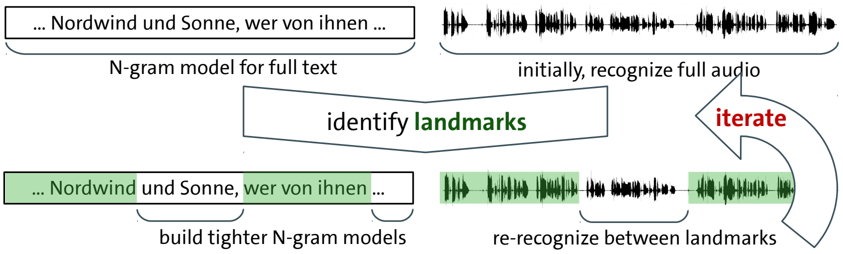

... Nordwind und Sonne, wer von ihnen ...
N-gram model for full text
initially, recognize full audio
identify landmarks
iterate
... Nordwind und Sonne, wer von ihnen ...
build tighter N-gram models
re-recognize between landmarks

Fig. 6: Audio alignment as repeated speech recognition

successively more restricted speech recognition: The main idea is to generate an n-gram model from the text to be aligned and use this for speech recognition on the provided audio. The speech recognition system generates a time-stamped word sequence. This sequence is matched against the original text and if a sequence of five or more words matches, the alignment for these words is kept as a landmark. The algorithm then recursively aligns the text and audio by splitting both audio and text between the landmarks and running the algorithm on each of these sub-ranges. The process stops once no new landmarks have been found. In this aspect we deviate from Katsamanis et al (2011), who use a fixed set of iterations.

Finally, if a word remains unaligned but both preceding and succeeding words have timings, we could infer the missing timing. However, we found that such words often mis-normalized (e. g. “*” (in German) is often spoken as “geboren” (*born*) but normalized as “Sternchen” (*asterisk*)) and should therefore not be aligned. Overall, in our implementation, we favor quality over coverage.

3.5 Phone Alignment

In principle we could have added phoneme timings as found by the alignment process but we opted against this given that the underlying speech recognition is not tuned towards identifying the pronunciation variant of a word or generating good timings for sub-word units. We instead derive phone-based alignments using MAUS (Schiel, 2004) which is specialized for this task.

MAUS works on comparatively small audio segments. Using the alignments already obtained by long audio alignment, we extract sentences from the text as well as the corresponding audio if the beginning and end of the sentence have timings, and the sentence contains only few and individual un-aligned words (with aligned words on both sides) that do not need normalization. The text is then converted to a phoneme-based representation using grapheme-to-phoneme conversion. Based on this representation and the audio, MAUS generates a phone alignment which is then re-added to the SWC XML (as a `ph` tag below the tokenization layer).

The phone alignments produced by MAUS can result in different word timings than the alignments produced by Sphinx. Adjusting one timing to match the other would require to know which one is correct, which we do not. We therefore provide both, the timings from alignment as part of the normalization layer and phoneme timings on the phoneme layer. Overall, we find that phoneme timings are generally more precise (measured at word boundaries) but the coverage of phone alignments in the corpus is lower.

3.6 Implementation

Our software suite is implemented in such a way that it allows for modular and distributed processing of the very large amounts of data.

All functionalities can be controlled using one reasonably simple commandline script. This main script handles the installation of the software (considerably lightening the deployment task) as well as delegation of downloading, tokenization, normalization and alignment to other software artifacts of the suite.

After downloading, further processing steps can be run per-article in different jobs. The alignment is especially CPU intensive and may need more than 10 times the audio length in CPU time for long audio files, making distribution over multiple machines almost mandatory. Jobs are simple shell scripts and can be handled by any clustering solution (such as `slurm` or `sge`) or may be run using a queueing mechanism built-in to the main script.

Steps are also decoupled and can be run individually consuming and producing valid SWC XML files that get gradually more richly annotated. This property may help in the future to integrate language-specific steps in a modular way.

A validation tool is available that tests both the adherence to the SWC XML schema as well as the character-level text correspondence with the original HTML.

4 Resulting Data Sets

We aligned the German, English and Dutch Spoken Wikipediae and make the results freely available. The resulting aligned data is annotated in the XML format described in Section 3.1 and uses the HTML rendering of the WikiText source received via the Wikipedia API as base. The underlying text characters (i. e. what remains after stripping all tags) contained in the HTML and SWC-XML tracks are identical. The text is tokenized and marked with sentence, paragraph and section boundaries. Each token is annotated with its normalization(s) and each normalization has its start and end timings if it was successfully aligned. For German and English we also provide segment-level annotations.

We perform our experiments on a heterogeneous collection of Linux-based compute nodes and workstations that use the built-in queueing mechanism of our main script. As the acoustic frontend is highly susceptible to DC noise in the recordings, we low-pass filter all audio which radically improve coverage for recordings on low-quality equipment. In the experimental results reported below, we include

Table 2: Overall size, alignment results and coverage for German, English and Dutch.

		German	English	Dutch
articles	total in category	1014	1339	3171
	downloaded and normalized	1011	1314	3073
	succesfully aligned	1010	1314	3073
hours	of audio	386 h	395 h	224 h
	of speech (after VAD)	360 h	364 h	204 h
	word aligned	249 h	182 h	21/79 h
	phonetically aligned	129 h	77 h	—
text	total characters	19.6 M	24.5 M	9.8 M
	total words	3.1 M	4.6 M	1.8 M
	words aligned	1.9 M	1.7 M	0.1/0.6 M
	sentences fully aligned	68 k	46 k	—
proportions	words w/ alignment	61 %	36 %	36 %
	speech in audio	93 %	92 %	91 %
	speech w/ alignment	69 %	50 %	39 %
	speech w/ phone al.	36 %	21 %	—

voice-activity detection results based on the WebRTC (Burnett et al, 2017) reference implementation.[12]

4.1 German Data

For German, as shown in Table 2, we aligned 1010 articles, containing 386 hours of audio. The audio contains about 360 h of speech (92 % of the audio, the remainder being a collection of 26 hours of silence ‘in the wild’). We count 3.1 million word tokens (in about 180 k word forms), out of which we align 1.9 M tokens for a total proportion of 61 % of the tokens and 69 % of the speech audio. The higher proportion of speech covered than tokens covered indicates that more often tokens are unspoken vs. unwritten speech is introduced.

In total, we align 249 hours of speech which, to our knowledge, is the largest corpus of freely-available aligned speech for German. In addition, we provide phonetic alignments for 129 h of speech, 36 % of the total audio, in 68 k fully-aligned sentences, which may be helpful for some analyses or training procedures.

Regarding the 288 aligned German readers, Figure 3 contains their respective alignment proportion (lower blue points). As can be seen in that figure, the proportion of audio aligned is not uniform across all speakers. In particular, there seem to be outlier speakers which were harder to recognize (maybe due to dialect, challenging recording conditions, or too lossy encoding). In principle, differences by speaker could also stem from different pausing behaviour by the speakers which would result in a higher proportion of silence that is not assigned to any aligned word (which we have not checked per-speaker).

For acoustic modelling, we used an iterative bootstrapping procedure that we adopted to counterbalance the limited-quality acoustic models that we started from:

[12] `https://github.com/wiseman/py-webrtcvad`, we report mean results of settings ‘1’ and ‘2’.

we first used the 2016 acoustic model from Voxforge which provides the best available freely licensed speech recognition models. We then built an acoustic model using SphinxTrain which we again use for aligning the full corpus. We then built new models based on this data as well as the Voxforge corpus and re-aligned. While our first model was only able to align 68h, this grew with every iteration reaching the numbers given above after 4 iterations.

Our software pipeline uses text normalization to account for numbers, dates, abbreviations, etc. We use MaryTTS which only supports a limited set of languages and does not support Dutch. We estimate the effect of normalization on a subset of the German corpus (10 % of the articles) for which we left out normalization prior to alignment. As a result, grapheme-to-phoneme conversation could not assign any meaningful phoneme sequence to such tokens, which reduces the alignment coverage (and potentially quality). We find that the alignment coverage only reduces by about 2 percentage points, indicating that normalization is not strictly required for large-coverage text-speech alignment. However, the effect on the proportion of fully aligned sentences is higher.

4.2 English Data

For English we aligned 1314 articles, containing 364 hours of speech (92 % of the audio, the remainder being silence) and 4.6 million tokens. We successfully aligned 182 hours of audio (50 % of the speech data) to their respective 1.7 M word tokens (36 % of the text data).

While the proportion of aligned audio may seem low, this results from frequent omissions of few words (particularly contractions "'s" and "'re") inbetween aligned portions. It would be easy to estimate reasonable timings for these words although these are not explicitly marked in our corpus. As a result of these frequent short omissions, the coverage of phonetic alignments is lower, at 21 % of the corpus.

In total, we align a slightly lower proportion of speech in total than for German, which may be due to the fact that we do not perform any re-training of acoustic models. For the alignment, we have used current off-the-shelf Sphinx-4 acoustic models (version 5.2 PTM) for US-English (despite the fact that the corpus contains many English dialects and also non-native speech).

4.3 Dutch Data

We include numbers for Dutch because free model training data and free acoustic models for Dutch are not readily available. However, this challenge limits the coverage of alignments. We use a model based on the IFA corpus (Son et al, 2001) as released by Voxforge in 2016. The model is based on only about 5 hours of speech, making the hundreds of hours of the Spoken Wikipedia a tempting source for speech material.

As a further restriction, the text normalization component of our alignment pipeline is not available for Dutch. Thus, at present, we are not using text normalization at all, precluding the alignment of numbers or abbreviations (and confusing the alignment in their surroundings). However, we estimate this effect to be small given the experiment on German data reported above.

Using the very limited model available and under the restrictions outlined above, we have been able to perform the alignment on 3073 articles (with a focus on shorter articles). Out of a total of 204 hours of speech we are able to align 21 hours in a first iteration. We have then used this data to re-train acoustic models for Dutch using standard SphinxTrain settings (and no knowledge of Dutch ourselves). We have used that model to re-align the corpus and reach a coverage of 79 h or 39 % of the speech.

This result indicates that despite the coarse approach (very limited acoustic model and no text normalization), large amounts of additional free speech material can be made available for research by exploiting the Spoken Wikipedia. The results given above increase the freely available (and freely licensed) aligned Dutch speech data 16-fold.

4.4 Verification of Alignment Quality

We have manually annotated the word-level boundaries for one German spoken article (Photodiode) containing 859 words in order to evaluate the quality of the word boundaries derived from automatic alignment. For every word token that was automatically aligned (coverage in this article was higher than average at 97.5 %), we computed the absolute deviation (in milliseconds) of the word boundary from that of the manual gold-standard alignment.

The deviations of the automatic alignment are presented as a histogram in Figure 7: we find that many (20 %) deviations are around 20-30 ms. The blue connected line shows the cumulative proportion of all word boundaries which deviate by no more than a certain timespan. We find that half the word boundaries found by the alignment lie within 40 ms of the gold standard, and 90 % within 150 ms, indicating a high alignment quality. Less than five percent of boundaries show errors larger than 200 ms.

The iterative bootstrapping procedure for acoustic models outlined in Subsection 4.1 above and executed for German and Dutch data could in principle result in overfitting our acoustic models to the data. Although this is unlikely with so many speakers in the corpus, we also checked the alignment quality of our German models on the Kiel Corpus of Read Speech (IPDS, 1994). We find that coverage is already very high with the first model and only marginally increases over the iterations, indicating that this corpus is easier to align than the Spoken Wikipedia. However, deviations from annotated phone boundaries in terms of RMSE continue to decrease with every iteration. This indicates that bootstrapping on the Spoken Wikipedia data does not overfit the model (likely due to the large size of the corpus) and leads to models that also perform better outside the training domain.

4.5 Error Analysis

Since a relevant portion of the data could not be aligned successfully, we performed a qualitative error analysis. We found several error types which we describe here.

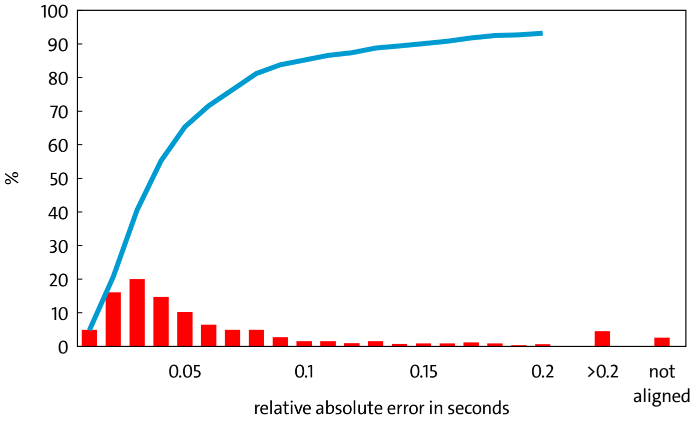


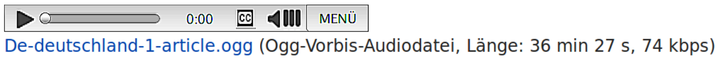
Fig. 7: Distribution of absolute alignment precision compared to manual alignment. Most words that can be aligned automatically have errors smaller than 100 ms.

Sometimes, the acoustic quality is simply too low for good recognition, either because of loud noise (which could potentially be filtered out) or distorting microphones. Accents such as Swiss German are also a problem for the alignment – most likely due to the acoustic models not fitting – as well as synthetic voices (which are used for some English articles). Some audio contains music, which can also not be aligned. For some (particularly Dutch) articles, only a part has been read at all (often the beginning of the article up to the first section). These problems cannot be easily solved.

As noted before, Wikipedia is semi-structured and there is no automatic check for structural consistency and completeness of meta data. For several articles, the wrong version ID of the article was stored. As version IDs are global (they need to withstand the renaming of articles), Wikipedia then simply serves a completely different article which does not match the expected textual reference at all. It should be noted though that our pipeline does not frequently align non-fitting text in these cases, demonstrating its robustness. Also, we were able to recover many of these errors and have since fixed them at the source.

Another class of errors stems from pronunciations which differ from the expected ones. The normalization fails in different ways (e. g. “Papst Pius XI” is spoken as “Papst Pius der Elfte”, *Papst Pius the eleventh*) and loan words such as “Engagement” do not match the expected pronunciation (because of errors in phonemisation or the speaker). For words not consisting of Latin characters such as Chinese names, we cannot generate a pronunciation at all. Finally, the text normalization to just one possible pronunciation is necessarily too narrow and it might be fruitful to input multiple alternative normalization options into the alignment process.

For English, we were especially interested in the low percentage of whole sentence alignments, which is only half that for German. It turns out that function words (e. g. *a, the, of, in,'s*) are often not aligned because they were not normalized to their reduced form. Such errors in the ASR-based alignment are often fixed by the phonetic alignment with MAUS which provides for phonotactic rules to generate reduced forms.



De-deutschland-1-article.ogg (Ogg-Vorbis-Audiodatei, Länge: 36 min 27 s, 74 kbps)

Fig. 8: The built-in interface for accessing Spoken Wikipedia articles on Wikipedia. Neither the menu nor the CC button reveil any helpful information.

5 The Usability of *Hyperlistening*

In this section, we exemplify the use of the newly time-aligned corpus for interaction research. We present an aural interface which uses the text-speech alignments described in the previous sections in order to allow *hyperlistening* to encyclopedic content: using the underlying textual and structural representation allows the user to navigate and leap based on meaningful structural units (sentences, paragraphs, sections), to search for word occurrences, or to follow links represented in the original written document. In contrast, the standard interface to the Spoken Wikipedia provides only for the linear consumption and crude time-based navigation (cmp. Figure 8), omitting all the positive aspects of hypermedia.

Using recorded audio instead of a text to speech system eliminates the potentially important variable of naturalness from the experimental setup. Mis-pronunciations and suboptimal prosody can hinder both understanding and interaction and thus severely distract the user. Naturalness ratings have been shown to degrade when listening to synthesized speech for an extended period (Pincus et al, 2015) although high-quality synthetic voices can be rated as more natural than amateur speech for short prompts (Georgila et al, 2012). We briefly describe our scenario and system and present the results of a usability study in which we compare two interaction modalities for hyperlistening. Our findings about the usability of hyperlistening can carry over to other scenarios with manually read material such as audio books or read news articles as well as scenarios where TTS is employed.

5.1 Hyperlistening

Wikipedia content (and all Wiki content for that matter) comes in the form of a strongly interlinked *hypertext*. Hypertext adds to traditional text the means for reading along a self-chosen reading path (i.e., non-linearily), called *hyperreading* (Zhang, 2006). We define *hyperlistening* as the aural consumption of read-out hypertext where interaction (and hence the listening path) can be controlled through graphical or voice-based interaction (or other means).

Wikipedia provides indices, extensive structural information, and – most importantly – associative links to enable hyperreading. A common strategy in hyperreading Wikipedia is *leaping* between sections of articles and between articles based on links or structure (Zhang, 2006). The recent advent of *find as you type* in most browsers has made *text search* a frequently used strategy to find information in web pages (Spalteholz et al, 2007). As also mentioned by Zhang (2006), a disadvantage of hyperreading is the possibility of getting lost due to the flexibility of what to read next. Getting lost may be of particular concern when hyperlistening, as speech is such an inherently

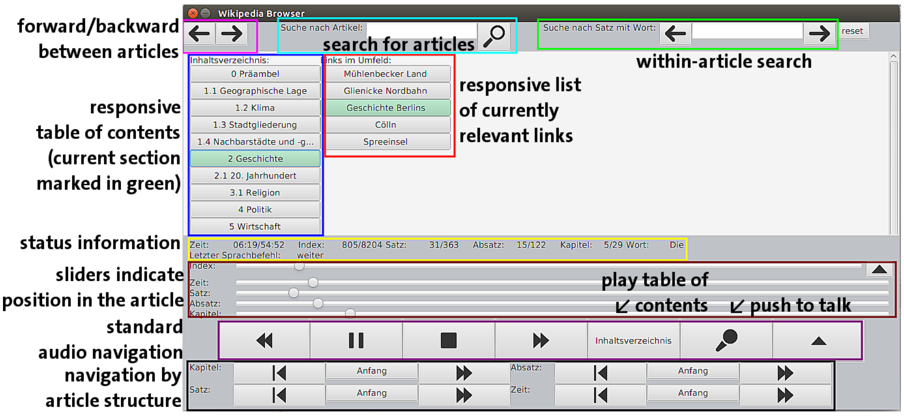


Fig. 9: The full application GUI for browsing the Spoken Wikipedia.

linear medium. Our experiment below focuses on whether participants are able to leap through speech without getting lost (too much), by assessing whether they are successful in navigating to key information in the article.

5.2 Prototype Implementation

We implemented a prototype for hyperlistening[13] that uses the original character data correspondence of the SWC XML created in the alignment process and the original HTML markup (cmp. Section 3.1) in order to infer the article structure headings and paragraphs as well as which tokens are part of hyperlinks to enable following such links during hyperlistening. We then add one graphical and one voice-based user interface that both allow to navigate based on the structural hierarchy of the article, the timing of all time-aligned words in the article, the sentence segmentation, and the hyperlinks contained in the article. We additionally synthesize the table of contents based on the observed article structure, as this is not spoken by the readers. We also outfitted our software with text-to-speech for articles that are not part of the corpus but left out this capability in the study presented below.

The *graphical user interface* consists of multiple parts that can each be hidden for experimentation. It is depicted in Figure 9 and offers various ways of accessing and leaping the structure of the article, as well as access to close-by links. In the experiments, the table-of-contents representation in the interface is delexicalized to avoid an unfair advantage in the graphical over the voice-based use-case.

The *voice user interface* for navigating spoken articles consists of speech activation, recognition and rule-based language understanding with the aim of offering similar functionality as the graphical interface. The user presses and holds the only button in the interface to activate speech recognition. When the button is released, we decode the recording using Google's freely available Speech API (Schalkwyk et al,

[13] Our implementation is available at `http://github.com/hainoon/wikipediareader`.

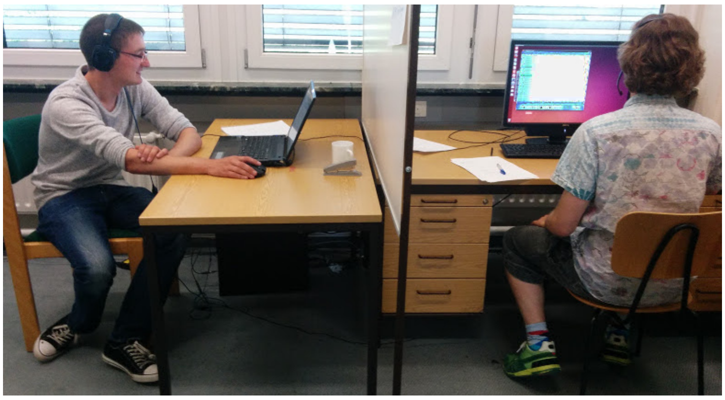


Fig. 10: Setup of the user study: the experiment participant (right side) and the experimenter/wizard (left side) are separated by a dividing wall.

2010)[14]. Language understanding makes use of all returned (n-best) hypotheses using a hierarchy of patterns-matching rules. Users may say (variations of) the following:

- "[show me the] [table of] contents",
- "next/previous chapter/section/paragraph/sentence",
- "[go back to the] beginning of the chapter/section/. . . " (or simply "repeat"),
- "[go to] chapter/section/subsection N",
- "*section name*" to go to the named section,
- "article *name*" to follow a link or search a named article.

5.3 User Study

We conducted a user study to gain insight into the overall usability of hyperlistening as well as the preferred modality for interaction. Our subjects' task was to navigate to information in the article that would help them answer simple factual questions in a limited amount of time. The information relevant to answer the questions were positioned anywhere in the article and sometimes required some simple combination. For our experiment we disabled the search and link-following in order to focus on structural navigation and leaping within the article. We compare three conditions:

GUI Users interacted using the graphical user interface as described above.

VUI Users interacted by speaking voice commands to the system. They were given a schema for possible commands.

Wizard-control As in the VUI setting, users interacted by speaking, but were instructed to use commands as they saw fit for the task (lead to believe that this was a 'better' system). In this condition, the experimenter followed the Wizard of Oz paradigm and navigated the article according to how the speech interface *should* act absent of recognition (and ensuing understanding) errors.

[14] `https://cloud.google.com/speech/`

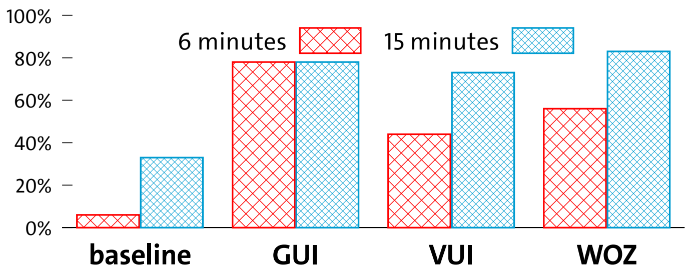


Fig. 11: Proportion of questions answered after 6/15 minutes for the experimental conditions and a non-interactive baseline.

As a **baseline** of non-interactive listening to a recording, we count all (randomly distributed) answers that appear in the test articles within the aloted browsing time. This baseline is optimistic as it assumes users to be able to correctly notice all answers without the ability to rewind/replay.

12 participants (normally sighted, not regular TTS or screen reader users) took part in the study. Each participant used the system in all three conditions and we balanced for ordering effects. Participants were given a choice of two articles for every condition so as to increase interest in the article in question. Participants where first allowed 2 minutes of ‘free browsing’ in the article. Afterwards, they were asked to use targeted navigation to answer three factual questions. The first 6 participants were allowed no more than 2 minutes for each question, the remaining 6 were allowed a total of 15 minutes with gentle reminders to move on after 5 minutes per question.

In all conditions, users wore a headset to listen to the system. The headset’s microphone was used only in the VUI condition, whereas the wizard directly heard the speaker and performed commands using the GUI from a separate computer. See Figure 10 for a picture of the setup. We asked the participants to fill out a questionnaire after the initial ’free browsing’ and after targeted navigation for each condition.

5.4 Results

Figure 11 shows the proportion of (fully or partially) correct answers under the three experimental conditions for the first group (6 minutes in total, 2 per question) and second group (15 minutes total), as well as the baseline. As can be seen, targeted navigation greatly improves factual information retrieval over linear listening in that 80 % of facts can be recovered hyperlistening for only 10 % – 26 % of the full article duration (6, resp. 15 minutes). We find that voice-based navigation profits from longer interactions, then reaching results on par with the GUI.

In the questionnaires, all conditions for hyperlistening were rated as ‘usable’ with a slight tendency towards spoken interaction (possibly because there is no modality change between output and input as commented by one user). Overall we find the effect of time to be higher than that of modality. Users tend to rate better when they have more time to interact, indicating that only 2 minutes per question result in stress, whereas 5 minutes are sufficient.

Regarding user behaviour, all participants interacted heavily (hyperlistened) in all conditions rather than listen linearly. In particular, they (a) navigate to sections, (b) skip ahead one section, paragraph or sentence, (c) go back one sentence when they notice that they found the desired information, or (d) pause playback. In voice-based interactions users often call the table of contents (before then calling for a section). VUI performance was somewhat restricted by errors which could be solved by better interface design (particularly wrt. push-to-talk) as shown in the wizarded condition.

Users reported that they easily stayed on top of things. This indicates that hyperlistening fits well with voice-based navigation and can hence be useful for eyes-free consumption of encyclopedic information.

Of course, the coverage of the Spoken Wikipedia is too limited at the moment for the spoken articles to replace TTS or screen-reading software completely. Hence, practical applications for aural Wikipedia browsing could combine text-to-speech and real-speech articles and correspondingly we have equipped our Wikipedia voice browser with this capability.

6 Conclusion and Discussion

We provide a novel time-aligned speech corpus of considerable size based on the Spoken Wikipedia as well as a process to automatically obtain more data over time for a variety of languages. To the best of our knowledge, we now provide the largest freely licensed speech corpora for German and Dutch, increasing the available free data by an order of magnitude. All data including the alignments and the acoustic speech recognition models that we produce from them is available under a Creative Commons license, as is the original material. We also provide our software tool for producing alignments under an open-source license to foster reproduction and extension of our work, for example to align other language versions of the Wikipedia.

We have certainly not yet maxed out the alignment coverage. For example, Tufiş et al (2014) report much better alignment coverage on conversational data from using speaker adaptation (although our endeavors into speaker adaptation did not show high yield). Neural network-based aligners have recently become available as well (McAuliffe et al, 2017) that promise even better performance. Correcting erroneous data in the Wikipedia and improving word normalization as well as using pronunciation alternatives could possibly also result in a higher alignment coverage.

So far, we have used the data for two purposes. We have trained acoustic models – in particular in order to bootstrap the alignment process – and these models show promise given that they are able to (a) bootstrap the alignment process, and (b) also result in qualitatively better alignments on a different corpus. We have also shown that the alignments improve aural access to encyclopedic content. In our experiment, we find that both modalities for interaction enable users to navigate articles and to find specific information much more quickly compared to the sequential presentation of the full article. We find that access to information is much faster when users are able to *hyperlisten* to spoken material rather than listen to it from beginning to end and we have shown that both conventional graphical browsing behaviour as well as spoken feedback for steering the browsing works reasonably well. Aural access to the Wikipedia is still limited. Screen-reading software provides an auditory rendition of the web and Wikipedia content is generally very well marked up semantically, yet screen-reading software has to be acquired and installed by the user. There is a specific webservice for the Wikipedia (the Pediaphon,[15] Bischoff (2007)) which downloads a TTS rendering of full articles. Pediaphon is based on MBROLA (Dutoit et al, 1996) which is not up to the state-of-the-art in the naturalness of the speech synthesized and likewise only allows for linear consumption.

Beyond speech recognition and aural access to encyclopedic material, the data that we provide could be a suitable source to create speech synthesis voices for those speakers who contributed large amounts of data (for German: 5 speakers with $>$ 2h fully aligned sentences and more yet unaligned data available). The present corpus, being non-literary factual speech, is likely well-suited for creating voices for present-day usages (e. g. in task-based dialog systems) as compared to other sources like LibriSpeech (Panayotov et al, 2015), which contain literary speech. A related project (Baumann, 2017) has recently analyzed the speakers in the SWC by their perceived speaking quality which further increases the value of our corpus.

A more recently proposed project, WikiSpeech (Andersson et al, 2016), aims to add open source text-to-speech capabilities to *MediaWiki*, the content management system behind Wikipedia (and many other wiki applications). That project proposes to bring tailored and multilingual text-to-speech that can be adapted to particular use-cases, improving the synthesis of encyclopedic entries. It is our hope that the Spoken Wikipedia Corpus will help to advance text-to-speech for encyclopedic content, given that it exhibits the prosodic and timing patterns that speakers use while reading out such texts.

Given the broad range of speakers involved in the corpora, we also intend to look at cross-speaker similarities and differences when reading out encyclopedic information. Finally, we want to use the resource to explore syntax-prosody correlations which is another reason why we focus on high-quality full-sentence alignments in our work.

Acknowledgments

We would like to thank all Wikipedia authors and speakers for creating this tremendous amount of data. We thank Florian Stegen for designing the initial version of the corpus extraction and alignment software that we built on. We also thank Marcel Rohde for running the experiment and collaborating on the initial workshop publication of the Spoken Wikipedia browser.

References

Ahn D, Jijkoun V, Mishne G, Müller K, de Rijke M, Schlobach S (2004) Using Wikipedia at the TREC QA track. In: Proceedings of the Thirteenth Text REtrieval

[15] http://www.pediaphon.org

Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, National Institute of Standards and Technology (NIST), vol Special Publication 500-261

Andersson J, Berlin S, Costa A, Berthelsen H, Lindgren H, Lindberg N, Beskow J, Edlund J, Gustafson J (2016) Wikispeech – enabling open source text-to-speech for Wikipedia. In: 9th ISCA Workshop on Speech Synthesis, Sunnyvale, CA, USA, pp 111–117, URL `http://ssw9.talp.cat/papers/ssw9_PS1-12_Andersson.pdf`

Baumann T (2017) Large-scale speaker ranking from crowdsourced pairwise listener ratings. In: Proceedings of Interspeech

Bischoff A (2007) The Pediaphon-speech interface to the free Wikipedia encyclopedia for mobile phones, PDA's and MP3-players. In: 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), IEEE, pp 575–579

Burnett D, Brandstetter T, Jennings C, Bergkvist A, Narayanan A, Aboba B (2017) WebRTC 1.0: Real-time communication between browsers. W3C working draft, W3C, https://www.w3.org/TR/2017/WD-webrtc-20170605/

Buscaldi D, Rosso P (2006) Mining knowledge from Wikipedia for the question answering task. In: Proceedings of the International Conference on Language Resources and Evaluation, pp 727–730

Chiarcos C, Dipper S, Götze M, Leser U, Lüdeling A, Ritz J, Stede M (2008) A flexible framework for integrating annotations from different tools and tag sets. Traitment automatique des langues 49:271–293

Dutoit T, Pagel V, Pierret N, Bataille F, Van der Vrecken O (1996) The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, IEEE, vol 3, pp 1393–1396

Ferres L, Sepúlveda JF (2011) Improving accessibility to mathematical formulas: The Wikipedia math accessor. In: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, ACM, New York, NY, USA, W4A '11, pp 25:1–25:9, DOI 10.1145/1969289.1969322, URL `http://doi.acm.org/10.1145/1969289.1969322`

Georgila K, Black A, Sagae K, Traum DR (2012) Practical evaluation of human and synthesized speech for virtual human dialogue systems. In: LREC, pp 3519–3526

Ghaddar A, Langlais P (2016) WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Grefenstette G (2016) Extracting weighted language lexicons from Wikipedia. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Hewlett D, Lacoste A, Jones L, Polosukhin I, Fandrianto A, Han J, Kelcey M, Berthelot D (2016) WikiReading: A novel large-scale language understanding task over Wikipedia. In: Proceedings of the 54th Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, pp 1535–1545, URL `http://www.aclweb.org/anthology/P16-1145`

Horn C, Manduca C, Kauchak D (2014) Learning a lexical simplifier using Wikipedia. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, pp 458–463, URL `http://www.aclweb.org/anthology/P14-2075`

Iftene A, Balahur-Dobrescu A (2008) Named entity relation mining using Wikipedia. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/

IPDS I (1994) The Kiel corpus of read speech. CD-ROM

Katsamanis A, Black M, Georgiou PG, Goldstein L, Narayanan S (2011) Sailalign: Robust long speech-text alignment. In: Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research

Köhn A, Stegen F, Baumann T (2016) Mining the Spoken Wikipedia for speech data and beyond. In: Proceedings of LREC, `urn:nbn:de:gbv:18-228-7-2209`

Laura Kassner VN, Strube M (2008) Acquiring a taxonomy from the German Wikipedia. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/

Lefever E, Hoste V, Cock MD (2012) Discovering missing Wikipedia inter-language links by means of cross-lingual word sense disambiguation. In: Chair) NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey

Max A, Wisniewski G (2010) Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta

McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M (2017) Montreal forced aligner: trainable text-speech alignment using kaldi. In: Proceedings of Interspeech

Nothman J, Murphy T, Curran JR (2009) Analysing Wikipedia and gold-standard corpora for NER training. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics, Athens, Greece, pp 612–620, URL `http://www.aclweb.org/anthology/E09-1070`

Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an asr corpus based on public domain audio books. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, pp 5206–5210

Pincus E, Georgila K, Traum D (2015) Which synthetic voice should i choose for an evocative task? In: 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, vol 105

Potthast M, Stein B, Gerling R (2008) Automatic vandalism detection in Wikipedia. In: European Conference on Information Retrieval, Springer, pp 663–668

Prabhakaran V, Rambow O (2016) A corpus of Wikipedia discussions: Over the years, with topic, power and gender labels. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Rohde M, Baumann T (2016) Navigating the Spoken Wikipedia. In: Proceedings of the Workshop on Spoken Language Processing for Assistive Technologies, San Francisco, USA, `urn:nbn:de:gbv:18-228-7-2290`

Schalkwyk J, Beeferman D, Beaufays F, Byrne B, Chelba C, Cohen M, Kamvar M, Strope B (2010) Your word is my command: Google search by voice: A case study. In: Advances in Speech Recognition, Springer, pp 61–90

Schiel F (2004) MAUS goes iterative. In: Proceedings of the LREC

Schröder M, Trouvain J (2003) The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6(3):365–377, DOI 10.1023/A:1025708916924

Son RJv, Binnenpoorte D, Heuvel Hvd, Pols LC (2001) The ifa corpus: a phonemically segmented dutch ”open source” speech database. In: Proceedings of Eurospeech, pp 2051–2054

Spalteholz L, Li KF, Livingston N (2007) Efficient navigation on the world wide web for the physically disabled. In: WEBIST (2), pp 321–327

Stegbauer C (2009) Wikipedia: Das Rätsel der Kooperation. Springer-Verlag

Strube M, Ponzetto SP (2006) WikiRelate! computing semantic relatedness using Wikipedia. In: AAAI, vol 6, pp 1419–1424

Suh B, Convertino G, Chi EH, Pirolli P (2009) The singularity is not near: Slowing growth of wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, ACM, New York, NY, USA, WikiSym ’09, pp 8:1–8:10, DOI 10.1145/1641309.1641322, URL `http://doi.acm.org/10.1145/1641309.1641322`

Tufiş D, Ion R, Ştefan Dumitrescu, Ştefănescu D (2014) Large smt data-sets extracted from Wikipedia. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), European Language Resources Association (ELRA), Reykjavik, Iceland

Walker W, Lamere P, Kwok P, Raj B, Singh R, Gouvea E, Wolf P, Woelfel J (2004) Sphinx-4: A flexible open source framework for speech recognition. Tech. rep., Sun Microsystems, Inc., Mountain View, USA

Wijaya DT, Nakashole N, Mitchell T (2015) ”a spousal relation begins with a deletion of engage and ends with an addition of divorce”: Learning state changing verbs from Wikipedia revision history. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, pp 518–523, URL `http://aclweb.org/anthology/D15-1059`

Yang D, Halfaker A, Kraut R, Hovy E (2016) Edit categories and editor role identification in Wikipedia. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik

M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Zesch T, Gurevych I (2007) Analysis of the Wikipedia category graph for NLP applications. In: Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007), pp 1–8

Zhang Y (2006) Wiki means more: hyperreading in Wikipedia. In: Proceedings of the Seventeenth conference on Hypertext and hypermedia, ACM, pp 23–26