Corpus Creation and Automatic Alignment of Historical Dutch Dialect Speech

Martijn Bentum[1], **Eric Sanders**[1], **Antal van den Bosch**[2], **Douwe Zeldenrust**[3], **Henk van den Heuvel**[1]
[1]Centre for Language Studies, Radboud University, Nijmegen, Netherlands
[2]Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands
[3]KNAW Meertens Institute, Amsterdam, The Netherlands
martijn.bentum@ru.nl, eric.sanders@ru.nl, a.p.j.vandenbosch@uu.nl,
douwe.zeldenrust@huygens.knaw.nl, henk.vandenheuvel@ru.nl

Abstract

The Dutch Dialect Database (also known as the 'Nederlandse Dialectenbank') contains dialectal variations of Dutch that were recorded all over the Netherlands in the second half of the twentieth century. A subset of these recordings of about 300 hours were enriched with manual orthographic transcriptions, using non-standard approximations of dialectal speech. In this paper we describe the creation of a corpus containing both the audio recordings and their corresponding transcriptions and focus on our method for aligning the recordings with the transcriptions and the metadata.

Keywords: dialectal speech, speech transcriptions, Dutch language variants, corpus creation

1. Introduction

The history of research on language variants is naturally tied to the history of national efforts, and efforts carried out transnationally in language areas, to record dialectal speech for posterity and future research. Both histories span more than a century. On the academic research side, descriptive dialectology and dialectometrics merged with language variation research and sociolinguistics. On the heritage side, recording and archiving dialect speech was (and often still is) the task of national committees and institutes, such as the Meertens Institute in the Netherlands, the Arni Magnusson Institute for Icelandic Studies in Iceland, or the Institute for Language and Folklore in Sweden.

These developments have been augmented with the advent of digitization and automated processing, which have allowed a broader access to speech resources and their metadata, and transcriptions, if available. Yet, historical dialectal speech[1] has remained a problematic data type for automatic speech recognition (Ghyselen et al., 2020; Miwa and Kai, 2023). Also, historical transcriptions, insofar as they are available and digitized, tend to pre-date or simply not follow standards such as the IPA, or canonical spelling rules. In this paper we show, using Dutch dialect data from the Meertens Institute as example, how non-standard orthographic transcriptions could still be aligned with dialectal speech, using state of the art speech processing methods. Our method should be generalizable to similar situations, for different language variations, assuming a general state-of-the-art speech recognizer such as Wav2vec 2.0 fine-tuned on a nearby standard language, such as in our case Dutch.

In a related study, Ghyselen and colleagues deal with the transcription hurdle in dialect corpus building (Ghyselen et al., 2020). The authors discuss the usefulness of ASR, respeaking, and forced alignment for a dialect corpus that is very similar to ours, viz. the Corpus of Southern Dutch Dialects (CSDD), covering West and East Flemish, Brabantian, and Limburgian. They address different transcription layers ranging from close to the pronunciations (for speech technology applications) to close to the standard language orthography (for NLP). With respect to forced alignment of the audio recordings to the transcriptions they concluded that forced alignment can be "very helpful to automatically refine the rough manual alignment of the transcription to the audio ... to a word-level alignment". This forced alignment, however, was based on a manually created orthographic transcription of the material. For the CSDD, as for our Meertens corpus, scans of handwritten and typed transcripts were also available. In the CSDD project it was decided to not use these, but to start from scratch with new handcrafted transcriptions. In our project we started from the OCR'd original transcripts as the starting point for alignment, avoiding the time-consuming step of creating new

[1]Throughout this paper we refer to *dialectal speech* and occasionally adopt the term *dialect*, while the current consensus is to speak of *(regional) language variants*, where variants are on an equal footing with what used to be called the *standard language*. In the Netherlands, Frisian, Limburgian and Low Saxon furthermore have formal statuses.

LREC-COLING 2024, pages 4021–4029

manual transcriptions. This implies that we have to accept OCR imperfections and original transcription inaccuracies and inconsistencies.

The aim of our method, speech recordings with time-aligned transcriptions, can benefit at least two possible goals: first, aligned data can be made accessible at fine-grained levels (words, phrases) in search engines, allowing users to find speech snippets containing their search terms. Second, the aligned data could be used to train or fine-tune a speech recognizer on the particular language variant at hand, which for less-resourced language variants such as Limburgian, Brabantish, and Low Saxon are typically non-existent as yet.

In this paper, we first review the history of the Meertens Institute's language variant recordings. We then describe our method of aligning audio to transcriptions, after which we evaluate the method. We describe the actual corpus that was derived from the base data of the Meertens Institute, and we offer some conclusions and points for future research.

2. The Meertens Dialect Recordings

The Meertens Institute[2], an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), was founded in 1930 under the name 'Dialectenbureau', 'Dialect Bureau'. It possesses numerous collections. One of these collections is the 'Banden Vrij Gesprek', 'Tapes [of] Free Conversation' (van Oostendorp, 2014). It consists of over 1,000 hours of audio recordings of spontaneous conversation, in a variety of Dutch language variants. The collection contains recordings made by researchers of the Meertens Institute and copies of recordings made by other persons or organizations as well (Rensink, 1962).

The Meertens Institute has pioneered recording audio since the 1930s. From 1952 onwards the institute started archiving audio recordings for various projects in a systematic way (Rensink, 1962). Jo Daan, head of the dialect department of the institute, was the main driving force behind this shift towards recording the sounds of the dialects, next to the traditional written coverage (Daan, 2000).

Often, recordings are of informants who worked on a farm. The conversations are typically about the work they did, like haying and harvesting (Rensink, 1962). Special festivities and public holidays, such as New Year's Eve and Easter, are also part of the conversation topics. There were conditions for the way the recordings should be made. For example, the researcher should not influence the informant with his or her own language variant (Rensink, 1962). In some cases, attempts were made to eliminate the researchers' influence entirely, by having language variant speakers talk to each other, while the researcher retreated outside of the field of view (Rensink, 1962). At the same time the definition of a spontaneous conversation was rather loose (van Oostendorp, 2014). Monologues and interviews with the researcher were included as well.

The collection also contains documentation. For each recording a summary and information about the speakers, place, date etc. were logged in notebooks. Next, transcripts were made for a substantial amount of recordings. The general idea was to make the information in the recordings more readily available. To achieve this, the quickest possible route – for the institute at the time – was taken (Van Haeringen and Meertens, 1964): the audio was not transcribed phonetically, but in spelling. At the same time the transcribers were asked to note down everything that stood out in pronunciation (Van Haeringen and Meertens, 1964). As a result, the transcripts vary greatly, not only in quality but also in structure. For instance, some are typed and contain abundant extra information, while others are just in plain written form.

Even though all this extra information was available, the accessibility of the collection remained limited. The material resided at the institute and researchers could for instance ask for a copy of a tape, but such requests were rare. By the turn of the century this changed radically. The institute started to digitize the audio collections and metadata was added to databases. In 2009 the audio of the free conversation collection and the metadata were made accessible via internet (Zeldenrust and van Oostendorp, 2013). The data were published on the website of the Meertens Institute and the collections was named 'Soundbites'. The website provided access to the recordings using the metadata. It also included a visual interface that was called the 'sprekende kaart', speaking map. The latter is a representation of the data using the geographic locations. Soundbites proved popular and in 2009 alone it generated 432,894 pageviews.

In the period that followed the website was improved and enhanced regularly (Zeldenrust, 2014). Scans of transcripts were added, as well as the scans of the notebooks (over 11,000 scans in total). The name changed to the 'Nederlandse Dialectenbank' (Dutch Dialect Database) and collections containing Dutch spoken abroad were added[3] (Zeldenrust, 2016). Next, data was made available via other web-based platforms such as CLARIN (Common Language Resources and Technology Infrastructure) and DANS (Data Archiving and Networked Services, also a KNAW institute).

[2] https://meertens.knaw.nl/

[3] https://ndb.meertens.knaw.nl/

2.1. DDD Materials

The DDD contains 1,976 audio recordings of Dutch language variants, most (1,942) were recorded in the Netherlands (some in Belgium, Germany, and the United States). Language variants recorded include Low Saxon (also known as Low German), Limburgian, Brabantic, Zeelandic, Hollandic, and West Frisian. Each recording is metadated with a Kloeke code [4], a unique identifier linked to a specific village, town or city (van den Heuvel et al., 2016). Figure 1 shows the distribution of audio recordings in the Netherlands. In total there are 976 hours of audio recordings. A typical recording lasts about 30 minutes, but duration ranges between 1 and 75 minutes.

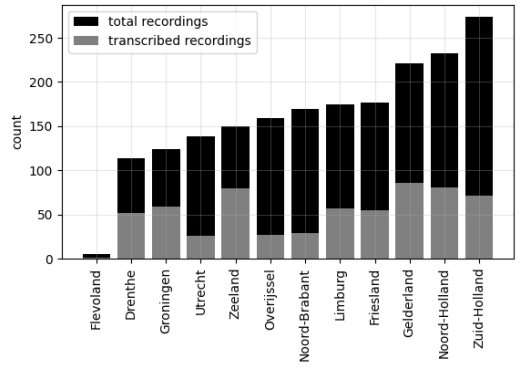

Figure 1: Number of recordings and transcribed recordings per province in the Netherlands

The collection of audio recordings started in earnest in the second half of the twentieth century with a peak during the 1970. Figure 2 shows the distribution of recordings over time.

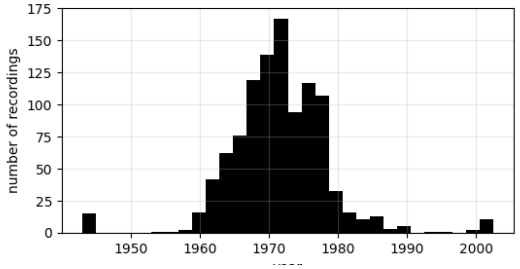

Figure 2: Number of recordings and transcribed recordings per province in the Netherlands

A number of recordings were manually transcribed. Most transcriptions (607) were created on a typewriter, while some (20) were handwritten. In Table 1 an overview is given of the transcribed materials. The materials were orthographically transcribed in an adjusted non-standard Dutch spelling to reflect the pronunciation of the recorded language variant. The result of a free spelling approach was that transcriptions contain many word types in part also because of the many contractions of multiple words with an apostrophe. This is reflected in the coverage of word tokens and types in Dutch Celex. Of the approximately four million word tokens only 55% are in Dutch Celex and of the almost 300,000 word types only 5% are in Dutch Celex.

Table 1: Overview of the manually transcribed recordings in the DDD. Twenty handwritten transcriptions were excluded from this overview, because the OCR was unreliable.

recordings	607
duration in hours	309
pages	11,751
sentences	452,629
word tokens	3,864,353
word types	263,174

3. Alignment of Audio and Transcriptions

3.1. Alignment on Recording Level

The data as it was delivered consisted of three separated parts: audio files, transcription files and one metadata file. There was no direct connection between the audio files and the transcription files (through the directory structure or the file naming), so we had to find the corresponding files ourselves. This was done by means of a Python script. All transcription files were searched. A part of the directory structure lead to the directory where the corresponding sound file was. In that directory were sound files named after the place where the recording took place, but the directory contained files from several recordings from different places. Another part (sub directory) of the transcription directory path was the recording ID, that could be used to look up the place name of the recording in the metadata file. Unfortunately, the spelling was not always the same as that of the sound file, so we had to look for the closest textual match between the the sound files in the corresponding directory and the place name to get the right sound file. For each recording we detected one audio file. The transcriptions consist often of several files. The original delivered transcription files are screenshots of the transcription. Later, the Meertens Institute delivered OCR'd versions of the transcriptions that were of very good quality.

By listening to the audio files and looking at the transcriptions it became clear that not all transcriptions and sound files that were brought together

[4] https://projecten.meertens.knaw.nl/mand/ECARTkartografiefie.html

in this way, fully corresponded. In a number of cases, part of the transcription was missing, and in other cases part of the audio was missing and a number of combinations were a complete mismatch. To find the (partial) mismatches, we looked at the number of times the automatic alignment (see section 3.2) did not detect a time stamp for the alignment. The recordings with the highest percentages of missing time stamps were checked manually. This way we could leave out the recordings for which the audio and transcription did not match.

3.2. Alignment on Sentence Level

The dialect recordings in the DDD can be an hour or longer in duration. The time alignment of long audio and text is a well studied problem (Wheatley et al., 1992; Moreno et al., 1998; Gao et al., 2009), however time aligning dialectal materials presents an extra challenge, because the transcriptions and speech do not match standard Dutch. See for example the excerpt below, with the dialect transcription (as present in the corpus), a standard Dutch transcription, and an English translation. The dialectal variants 'tied' (time), 'meulenears' (millers) and 'touwtj'an' (string to) will not be part of a standard lexicon of Dutch and the ASR-system might therefore also struggle with the pronunciation of the recorded speech. However, there is still a lot of (be it imperfect) correspondence between the dialectal and standard version of the transcription.

dialectal: in de tied van de meulenaers dan bonde we 'n touwtj'an de meulenaer
standard: in de tijd van de molenaars dan bonden we een touwtje aan de molenaar
translation: in the time of the millers then we bound a string to the miller

As a benchmark for forced alignment we utilized the web-based Maus forced aligner[5] (Schiel, 1999; Kisler et al., 2017). This aligner maps the sounds in the audio to a phone-based representation of the given transcription by using a hybrid approach consisting of statistical classification of the signal (HMM) and probabilistic rule based components (derived from corpus statistics). Possible pronunciation variants and out of vocabulary words are taken into account. This renders the method applicable to read speech as well as to spontaneous speech.
We compared the Maus forced aligner to our new alternative approach based on Wav2vec 2.0 (Baevski et al., 2020). End-to-end models such as Wav2vec 2.0 can generate usable transcriptions without the application of a lexicon or language model. We can thus generate an automatic transcription by applying a Wav2vec 2.0 model fine-tuned for standard Dutch on the speech recordings. Subsequently, with the aid of the Needlemann-Wunsch algorithm (Needleman and Wunsch, 1970), a similarity search method developed to align amino acid sequences (e.g. DNA), it is possible to align this automatic transcription to the dialectal transcription based on the correspondence between the dialectal and standard transcriptions. We can then map the timestamps from the automatic transcription to the dialectal transcription and finally align the dialectal transcription to the audio recordings.

We will compare this new alignment technique with the Maus forced aligner. We also validated the new alignment technique on a subset of the speech materials in the DDD by manually annotating the alignment quality. Furthermore, we validated the new alignment technique on corpus materials from the Spoken Dutch Corpus (Oostdijk, 2002). This corpus contains Netherlandic and Flemish Dutch speech recordings for which all transcriptions and alignments were manually checked. These materials allowed us to automatically validate alignment and test the performance on a larger data set, estimate the mismatch of timestamps and study the influence of mismatching transcriptions. It also informs us about the maximum achievable quality of alignment between recording and transcription under ideal conditions.

3.2.1. Materials

For the alignment experiments we utilized materials from two corpora: the DDD and the Spoken Dutch Corpus (Oostdijk, 2002). From the DDD we selected all manually transcribed recordings in the *Nedersaksisch* region (part of the eastern Dutch provinces Gelderland and Overijssel) with approximately 30 hours of materials. For each recording with a transcription available, we collected the transcription in a text file. The text was cleaned by removing upper case, interpunction (except apostrophe) and references to the speaker (e.g. 'A:' at the start of a transcription line).

For the Spoken Dutch Corpus we created two data sets. For one data set we selected all Netherlandic Dutch recordings, approximately 500 hours of speech materials. For the other data set we randomly selected 10% of the recordings from the first data set. We cleaned the text by removing upper case, interpunction (except apostrophe), special word codes (i.e. * with a letter indicating a special word status) and removing other special codes indicating pronunciation phenomena, such as laughter and unintelligible speech (e.g. ggg or xxx).

[5] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

3.2.2. Alignment with Wav2vec 2.0

For each audio file in the selected materials, we applied a Wav2vec 2.0 model fine-tuned for standard Dutch[6] and extracted both the transcribed text and the timestamps for each character.
Next, we applied the Needleman-Wunsch (NW) algorithm with a score of 1 assigned to match, mismatch and gap. These scores indicate a bonus of 1 to matching characters, a penalty of 1 to mismatching characters and a penalty of 1 to gaps (the - characters in the example below). Based on these scores the NW algorithm finds an optimal alignment by maximizing the overall score. We did not do any parameter tuning since the default setting resulted in satisfactory performance. We applied the NW algorithm to each text file with the cleaned version of the original transcription and the corresponding Wav2vec 2.0 transcription, resulting in an aligned version, see the example below, with the O line containing the original transcription and the A line the Wav2vec 2.0 transcription.

```
O: sloapkoamer joa hen'ezakt b-ezunder
A: slaapk-amer j-a--en- za-t bijzonder
```

The unit we seek to align is the sentence. For the DDD, the sentence was defined as the words on a single line in the original transcription document (on average 11.87 words per sentence). For the Spoken Dutch Corpus, we defined a sentence as the words contained by a single transcription segment (on average 7.33 words per sentence). The sentences in the original transcription can be aligned with the audio by a look up of the timestamp of the Wav2vec 2.0 transcriptions character corresponding to the start and end of the sentence (the starting *s* and final *r* respectively in the above example).

3.2.3. Manual validation

The alignment of the dialect speech and transcriptions was manually checked with a web annotation tool (based on the Django framework) built for this purpose. The audio files were split into sentences based on the timestamps from the alignment technique. We presented the speech recording in combination with the corresponding original transcription to the annotator. The annotator listened to the audio and checked whether it matched the transcription and indicated one of six possible options, the labels *good* or *bad* for sentences that were or were not correctly aligned respectively. The annotators also assigned the label *good* if they heard at most one syllable more at the beginning or end of a sentence compared to the transcription. We also used the labels *start match, end match, middle match*, and *middle mismatch* to indicate partial alignment.

[6] https://huggingface.co/FremyCompany/xls-r-2b-nl-v2_lm-5gram-os

3.2.4. Automatic validation

To validate the alignment technique on the Spoken Dutch Corpus without manual annotations, we utilized the timestamps of the original transcriptions of the Spoken Dutch Corpus to automatically label the match between the transcription generated by the Wav2Vec 2.0 ASR and the original transcription. We used the same set of labels, except for middle mismatch, which we could not reliably automatically assign (however this label also almost never occurred in the manually labeled set). We assigned the label *good* if the aligned transcription matches the original transcription and the start and end times matched within a margin of 0.5 seconds. For *start match* and *end match* the transcription should match and the start or end time (respectively) should be within the margin of 0.5 seconds. For *middle match*, the transcription should match and the start and end time should overlap with the original transcription. The label *bad* was assigned when the transcription did not match or the start and end times did not overlap with the original transcription.

The automatic alignment labelling of the Spoken Dutch Corpus does not directly translate to the manual labelling of the language variant materials. The validation with the Spoken Dutch Corpus is intended to provide further insight into the quality of the alignment technique and not to directly compare with the results we obtained with the language variant materials. Also, the availability of a ground truth of time alignments in the Spoken Dutch Corpus allowed us to further investigate the alignment technique.

We also investigated the sensitivity of this alignment approach to mismatch in spelling, since the orthographic transcriptions of the DDD do not adhere to standard Dutch spelling. Based on the materials from the Spoken Dutch Corpus, we investigated the influence of mismatches between the original transcript and the speech recording on alignment quality. For this purpose we created altered versions of the transcriptions from the Spoken Dutch Corpus. We altered the transcriptions by randomly reassigning a certain percentage of characters to other characters. We created randomized versions of the transcripts with 2, 4, 8, 32 and 64% of the characters replaced. To limit the computational load of this experiment we applied this test to 10% of the corpus ($\sim$ 50 hours).

4. Tests and Evaluation

4.1. Maus versus Wav2vec 2.0

We compared the forced alignment results from two different forced aligners, namely Maus and our Wav2vec 2.0 based approach. For Maus we used the 'Pipeline with ASR' with the option 'GP2->CHUNKER->MAUS'. We applied the forced aligners on a sample of sentences from four one hour recordings from the DDD. Subsequently, four human annotators labelled the alignment of a sample of sentences from each recording. Each sentence was labelled by two annotators. Table 2 shows the counts for the alignment labels. The first count in each cell corresponds to the Maus forced aligner and the second count to the Wav2vec 2.0 based aligner.

For both the Rekken en Lievelde recordings the Maus forced aligner could not correctly align any transcription with the audio. This is likely due to poor chunking, a step that the Maus forced alignment pipeline applies to longer audio files, because they become prohibitively computationally expensive to force align and are therefore chunked into smaller sections. For the Rekken and Lievelde recording, we needed to lower the 'Minimum anchor length' of the Maus chunker, which adversely affected the chunk quality. The Wav2vec 2.0 based aligner does not have a chunking step.

Since the Maus alignment for Rekken and Lievelde were completely misaligned we excluded these recordings from further analysis, because the annotated labels showed no variability (i.e., the Wav2vec 2.0 based aligner outperformed the Maus aligner). For the Keyenburg and Haarlo recordings, we first computed the inter-rater reliability for the whole label set (good, start match, middle match, end match, bad): Kappa = .57, this is a fair to good agreement, with 72% overlap in the assigned labels. In addition, we computed the inter-rater reliability for a reduced label set, comparing bad labels versus all other labels (i.e. good, start match, middle match and end match) combined: Kappa = .87, this is an excellent agreement with 96% overlap in the assigned labels.

We compared the alignment performance of the Maus and the Wav2vec 2.0 aligner based on the Keyenburg and Haarlo recordings. For the comparison the bad labels versus all other labels and randomly selected one annotation per sentence (each sentence was annotated by two annotators) for each annotated sentence. The Maus aligner scored 92 good and 7 bad alignment labels, the Wav2vec 2.0 based aligner scored 79 good and 21 bad alignment labels. The results of the Fisher's exact test (3.49, $p < 0.01$) indicate that the Maus aligner performed significantly better compared to the Wav2vec 2.0 aligner on the Keyenburg and Haarlo recordings. The Maus aligner is the preferred option when robust chunking is possible, otherwise the Wav2vec 2.0 aligner is the better choice. Since half our sample could not be chunked correctly we decided for our corpus collection to align the recordings and transcriptions with the Wav2vec 2.0 based aligner. The following sections detail further validation tests we performed.

4.2. Validation of Wav2vec 2.0 based forced alignment

The alignment label counts and percentages can be found in Table 3, for both the dialectal Nedersaksisch materials (from the DDD) and the Netherlandic materials from the Spoken Dutch corpus.

The dialectal materials were manually annotated by two annotators trained in transcribing speech, who annotated 16,460 unique sentences. Two audio files did not match with their respective transcriptions, due to an error in the metadata of the corpus. The 935 sentences associated with these two files (all with the alignment label *bad*) were removed from the analysis.

A subset of the sentences were annotated by both annotators. We computed the inter-annotator agreement for the complete labels set, Kappa = .234 (95% confidence interval: 0.222-0.245). This is a fair alignment, with a 40% overlap in the assigned labels. In addition, we checked whether the annotators matched in their judgement between *bad* versus the combined set of other labels and computed the inter-annotator agreement, Kappa = .912 (95% confidence interval: 0.898-0.926), with a 97.95% overlap in the assigned labels, confirming that the annotators were consistent in distinguishing between well aligned and badly aligned transcriptions.

For both the dialectal materials and the materials from the Spoken Dutch Corpus, most sentences ($\sim$ 90%) were at least partially matched or better. About a third of the sentences did show a mismatch between the start time, end time or both.

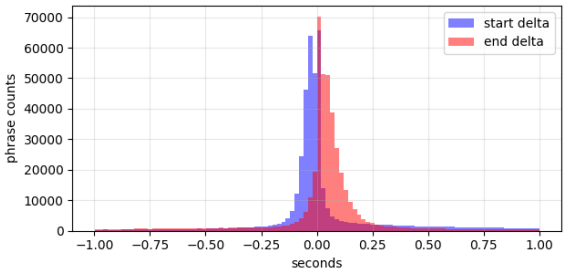

Figure 3: Misalignment in seconds of the start times (blue) and end times (red) between original and newly aligned sentences. Based on the materials from the Spoken Dutch Corpus.

The sentences from the Spoken Dutch Corpus were automatically labelled (see Section 3.2.4).

Table 2: Forced aligner comparison. Aligment label counts for a sample of sentences from four recordings from the Dutch Dialect Database. Each sentence was labelled by two annotators. The counts in each cell are for the Maus, and Wav2vec 2.0 forced aligners respectively.

Label	Rekken M | W	Keyenburg M | W	Lievelde M | W	Haarlo M | W
good	0 | 55	122 | 105	0 | 96	112 | 88
start match	1 | 28	28 | 26	0 | 24	27 | 34
end match	0 | 32	20 | 21	0 | 20	18 | 22
middle match	0 | 24	12 | 10	0 | 21	14 | 13
bad	201 | 60	16 | 36	200 | 38	29 | 43

Table 3: Aligment label counts and (percentages) for the DDD and the Spoken Dutch Corpus materials.

Label	DDD	Spoken Dutch Corpus
good	8,184 (52.71%)	302,242 (61.77%)
start match	2,069 (13.33%)	32,058 (6.55%)
end match	1,383 (8.9%)	48,711 (9.95%)
middle match	1,893 (12.19%)	71,921 (14.7%)
middle mismatch	327 (2.11%)	
bad	1,670 (10.76%)	34,390 (7.03%)

The assignment of the labels is dependent on the delta value used (i.e. the allowed mismatch in seconds between the start and end timestamp). For the results in Table 3 the automatic labelling was done with a delta of 0.5 seconds. To also give insight in the timing differences between the original sentences and the newly aligned sentences we plotted the distribution of differences in start and end times between original and newly aligned sentences (see Figure 3).

4.3. Transcription errors

We tested the robustness of the alignment technique by artificially introducing transcription errors in the 10% subset of the Spoken Dutch corpus materials. Figure 4 displays the percentage of sentences with a specific alignment label as a function of the percentage of characters substituted by a randomly chosen different character. The influence of the transcription errors is most pronounced in the *bad* and *middle match* category, and to a lesser extent in the *good* category. The introduction of transcription errors does not seem to have much effect on the *start* and *end match* categories. The alignment technique appears to be fairly robust against noisy transcriptions; even with more than half of the characters assigned to a random other character, only 30% of the sentences is completely misaligned.

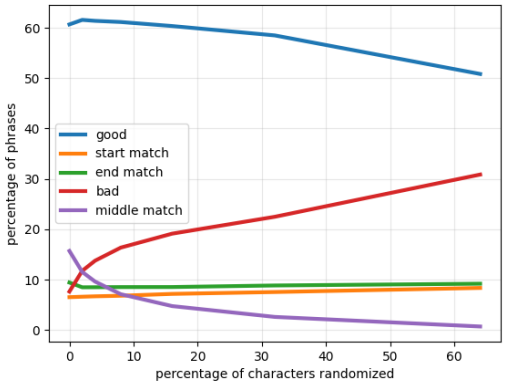

Figure 4: Alignment quality as a function of percentage of characters in transcription that were randomly replaced. Based on the 10% subset of the Spoken Dutch Corpus materials

5. Corpus Compilation

The audio, transcription, metadata and alignments are compiled into a structured corpus. Each recording is stored in a separate directory. One such directory contains the audio file in mp3 format (the original format in which it was delivered), the screenshots of the transcription pages, a json file with all transcriptions after OCR, metadata and alignment information and Praat [7] textgrids. The names of the file are the original ID of the recordings as they were in the metadata file. This was done for backwards compatibility. The textgrids contain three tiers:

- manual transcription: Contains intervals with OCR lines for which the Wav2vec 2.0 based alignment method found timestamps.

[7] http://praat.org

- overlapping transcription: Contains intervals with timestamps that overlap with other intervals.

- not aligned: Contains intervals for OCR lines for which time stamps were not found. The time stamps were taken from neighboring intervals with timestamps.

6. Discussion and Conclusion

We described the steps taken to convert a collection of speech recordings and OCR'd transcriptions to an aligned corpus. The starting situation is typical for research on heritage speech data. The base materials (speech recordings on magnetic tape and transcriptions on paper) were digitized, but for automated processing two problems are compounding: first, the speech is non-standard and recognized less accurately than present-day standard Dutch, and second, transcriptions use a non-standard Dutch spelling to approximate dialectal pronunciation and non-standard dialect words that are not part of the standard present-day Dutch lexicon.

Our proposed solution to align the dialectal speech recordings and transcriptions, is to make use of an end-to-end speech recognizer, Wav2vec 2.0, and the Needlemann-Wunsch string alignment algorithm. We showed that this approach is able to align about 50% of sentences correctly and about 90% of sentences at least partially correct for a subset of DDD materials. Using the Spoken Dutch Corpus as a manually checked gold standard, we were able to demonstrate that our method is quite robust; with half of the characters randomized, our system still only misaligned about 30% of all sentences. The proposed alignment method could be applied in similar cases, provided that a Wav2vec 2.0 model can be fine-tuned for a language variant (which would usually be a variant considered more standard, for which ample data is available) that is close enough to the language variant at hand.

The aligned materials in the DDD are suitable in principle for word-based search engines to produce sound snippets containing the query words. However, it is important to note that only a small portion of the materials in the DDD was manually checked and the aligned speech recordings and transcriptions should not be treated as a gold standard.

The materials in DDD could also be used as training data for a new dialectal speech recognizer, however, this would be non-trivial, because the non-standard Dutch spelling used in the dialectal transcriptions is highly variable. The transcriptions were made by many annotators over a long period of time and it is much harder to spell consistently when not adhering to a spelling standard. To be able to use the materials in the DDD as training data for a dialectal speech recognizer it might be necessary to add a transcription tier with standard Dutch spelling or a standardized dialectal spelling. The corpus with the audio-transcription alignments created with this alignment approach will be distributed via the Meertens Institute and will be available as of January 2024.

7. Bibliographical References

Harald Baayen and Richard Piepenbrock and Leon Gulikers. 1996. *Celex2*. [link].

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Johanna Catharina Daan. 2000. *Geschiedenis van de dialectgeografie in het Nederlandse taalgebied: Rondom Kloeke en het Dialectenbureau*. Koninklijke Nederlandse Akademie van Wetenschappen.

Jie Gao, Qingwei Zhao, and Yonghong Yan. 2009. Online detecting end times of spoken utterances for synchronization of live speech and its transcripts. In *Tenth Annual Conference of the International Speech Communication Association*.

Anne-Sophie Ghyselen, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen, and Arjan van Hessen. 2020. Clearing the transcription hurdle in dialect corpus building : the corpus of southern dutch dialects as case-study. *Frontiers in Artificial Intelligence*, 3:10:1–10:17.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Shogo Miwa and Atsuhiko Kai. 2023. Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR. In *Proc. INTERSPEECH 2023*, pages 4928–4932.

Pedro J Moreno, Christopher F Joerg, Jean-Manuel Van Thong, and Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *ICSLP*, volume 98, pages 2711–2714.

Annie Murray and Jared Wiercinski. 2014. A design methodology for web-based sound archives. *DHQ: Digital Humanities Quarterly*, 8(2).

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Nelleke Oostdijk. 2002. The design of the Spoken Dutch Corpus. pages 105–112.

W.G. Rensink. 1962. Dialecten op band. *Taal En Tongval*, 14:184–196.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech.

Henk van den Heuvel, Eric Sanders, and Nicoline van der Sijs. 2016. Curation of Dutch regional dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3249–3255, Portorož, Slovenia. European Language Resources Association (ELRA).

C. B. Van Haeringen and P. J. Meertens. 1964. Verslag van de dialectencommissie der koninklijke akademie van wetenschappen te amsterdam over 1963.

Marc van Oostendorp. 2014. Phonological and phonetic databases at the meertens institute.

Barbara Wheatley, George Doddinton, Charles Hemphill, John Godfrey, Edward Holliman, Jane McDaniel, and Drew Fisher. 1992. Robust automatic time alignment of orthographic transcriptions with unconstrained speech. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 533–536. IEEE.

Douwe Zeldenrust. 2014. Access to data: The soundbites collection of the meertens institute and a flexible approach to the curation and dissemination of humanities digital resources. In *Digital Humanities Benelux Conference 2014*.

Douwe Zeldenrust. 2016. The creation, curation and dissemination of humanities digital resources: The dutch dialect database.

Douwe Zeldenrust and Marc van Oostendorp. 2013. Combining tailor made research solutions with big infrastructures: the speaking map of the netherlands. In *Digital Humanities Conference 2013, Proceedings.* University of Nebraska.