Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1009–1015
Marseille, 20-25 June 2022

A Speech Recognizer for Frisian/Dutch Council Meetings

Martijn Bentum[1], Louis ten Bosch[1], Henk van den Heuvel[1], Simone Wills[1], Domenique van der Niet[2], Jelske Dijkstra[3], Hans Van de Velde[3,4]
[1]CLS/CLST, Radboud University; [2]Humain'r; [3]Fryske Akademy; [4]Utrecht University
{martijn.bentum, louis.tenbosch, henk.vandenheuvel, simone.wills}@ru.nl, domenique@humainr.com,
{jdijkstra,hvandevelde}@fryske-akademy.nl

Abstract

We developed a bilingual Frisian/Dutch speech recognizer for council meetings in Fryslân (the Netherlands). During these meetings both Frisian and Dutch are spoken, and code switching between both languages shows up frequently. The new speech recognizer is based on an existing speech recognizer for Frisian and Dutch named *FAME!*, which was trained and tested on historical radio broadcasts. Adapting a speech recognizer for the council meeting domain is challenging because of acoustic background noise, speaker overlap and the jargon typically used in council meetings. To train the new recognizer, we used the radio broadcast materials utilized for the development of the *FAME!* recognizer and added newly created manually transcribed audio recordings of council meetings from eleven Frisian municipalities, the Frisian provincial council and the Frisian water board. The council meeting recordings consist of 49 hours of speech, with 26 hours of Frisian speech and 23 hours of Dutch speech. Furthermore, from the same sources, we obtained texts in the domain of council meetings containing 11 million words; 1.1 million Frisian words and 9.9 million Dutch words. We describe the methods used to train the new recognizer, report the observed word error rates, and perform an error analysis on remaining errors.

Keywords: ASR, Code Switching, Frisian, Dutch, Domain Adaptation

1. Introduction

We developed a new bilingual Frisian/Dutch automatic speech recognizer (ASR) for Frisian council meetings. During these council meetings both the Frisian and the Dutch language are spoken. We build on an existing Frisian/Dutch ASR system, entitled FAME! (Yılmaz et al., 2018), which was trained and tested on radio broadcasts. This FAME! speech recognizer was adapted for the domain of council meetings.

Adapting the ASR system for the Frisian council meeting domain is challenging because of the noisy background, the use of jargon and multiple overlapping speakers. During these meetings, long and complex words are used, typical for civil servants and government style documents, but not part of a general-purpose lexicon. For example, *ambtenaarûndersteuning* 'civil servant support' is a compound consisting of the stems *ambtenaar* 'civil servant' and *ûndersteun* 'support', or *kostprijsberekkening* 'cost price calculation' is a compound consisting of the stems *kost* 'cost', *prijs* 'price' and *berekken* 'calculate'.

For this project an automatic speech recognizer was built that could deal with Frisian and Dutch as well as with code switching between the two languages as is often encountered in Frisian. We trained a new speech recognizer based on the FAME! radio broadcast materials combined with new speech recordings. The FAME! materials (https://fame.ruhosting.nl/) contain more than 3,000 hours of Frisian and Dutch radio broadcasts from the Omrop Fryslân (Frisian Broadcast) covering the period 1950–2000. The new speech material consists of audio files with transcriptions of council meetings for training acoustic models (AM) and language models (LM) and of textual reports of council meetings for further language model training.

The new speech recognizer was created as part of a project by the Fryske Akademy to develop a subtitling service for council meetings in Frisian and Dutch, following the legal obligation to make the online recordings of public council meeting accessible for the deaf and hard of hearing. Until now, such services were only available in Dutch. The project was financed by the Province of Fryslân, the Frisian water board, and eleven Frisian municipalities. The new Frisian ASR is intended for a subtitling service deployed by the company Humain'r. The ASR system provides initial transcriptions of the meetings, which are monitored on quality standards before being converted, optimized and distributed in several subtitling formats via cloud API.

In this paper we study the steps needed to adapt an ASR system for the low resourced language Frisian to a new domain (council meetings). We test the influence of automatically added speaker labels and the addition of Dutch materials to recognition accuracy and report a detailed analysis of superficial errors to provide insight about the impact of these type of errors. Lastly, we present a new dataset with manually transcribed Frisian and Dutch speech that extends the previous FAME! dataset.

In the next sections we explain the method and data used for training the new ASR system, the results, and an error analysis. At the end of the report we present our conclusions and suggestions for further improvement.

2. Method

2.1 Data

We used the manually transcribed audio recordings in the FAME! radio broadcast corpus (Yılmaz, Van den Heuvel, & Van Leeuwen, 2018). The recordings contain approximately 11 hours of speech, with 8 hours of Frisian and 3 hours of Dutch. The annotations include speaker labels, which were used in the acoustic model training.

In addition, we used newly created manually transcribed audio recordings of council meetings from several Frisian municipalities (see below). The audio recordings consist of 49 hours of speech, with 26 hours of Frisian (281 thousand words) and 23 hours of Dutch speech (287 thousand words). Speaker labels were not part of the transcriptions.

The manual transcriptions were created by the company Humain'r for the purpose of training a multilingual ASR system. Humain'r extended their pre-existing Voice Technology Solution platform with multilingual capabilities and developed a new training program for the human transcribers for the task of transcribing Frisian/Dutch multilingual speech materials. This approach resulted in high quality transcriptions containing language and dialect labels (various Frisian and mixed Frisian-Dutch varieties, labels for code-switching and labels for frisisms in Dutch and dutchisms in Frisian. The resulting set of manually transcribed Frisian/Dutch speech represents a four-fold increase compared to manually transcribed materials available in the FAME! corpus.

Furthermore, we used the manually transcribed speech recordings in the Spoken Dutch Corpus (Oostdijk, 2002). These materials consist of approximately 750 hours of Dutch and Flemish speech. The annotations include speaker labels. These materials were used to train a diarization model (the data contain approximately 4,200 speakers) and to augment the Dutch training materials for the acoustic model training.

We also extended the text materials in the domain of council meetings. We collected texts such as council meeting minutes and council policy documents. The texts contain 11 million words of which 1.1 million Frisian words and 9.9 million Dutch words. Texts and audio recordings were provided by Provinsje Fryslân and Wetterskip Fryslân (Frisian water board) and the municipalities of Achtkarspelen, Dantumadiel, Fryske Marren, Heerenveen, Leeuwarden, Noardeast-Fryslân, Opsterland, Sudwest-Fryslân, Tytsjerksteradiel and Waadhoeke.

In addition to the new text materials, we used a pre-trained trigram model, which is part of the FAME! speech corpus (see Yılmaz, Van den Heuvel, & Van Leeuwen, 2018 for more details).

2.2 Training

The Frisian recognizer we developed is bilingual, recognizing both Dutch and Frisian. To accomplish this, we employed a language tag system similar to Yılmaz, Van den Heuvel, & Van Leeuwen (2018). This entails that each word and phone is language tagged, to keep the languages completely separate. The tag consists of a hyphen with a language id (-fr for Frisian and -nl for Dutch) appended to each word (i.e. orthographic form) and each phone thereof. In order to achieve this, all words in the new text materials needed to be classified according to their language. The council text materials contain both Dutch and Frisian words. To provide a language tag, we trained and applied a language classifier to tag the language of each word in the council material texts (see Appendix A for performance metrics).

2.2.1 Training, development, and test sets

We split the materials from the newly collected council meeting recordings into a training set, a development set, and a test set (respectively 80%, 10%, 10%). For the development and test set we selected sections of 15 minutes of consecutive materials from different meetings to ensure the tests were not biased to a specific meeting and to be able to demo the recognizer on longer stretches of held out data. To the training set, we added the training materials from the FAME! corpus. In addition, we used the test materials from the FAME! corpus separately to compare decoding results on the test set for the council meetings and the FAME! corpus.

2.2.2 Lexicon

For the decoding lexicon we took the lexicon as provided in the FAME! corpus as a starting point. We first excluded approximately 12 thousand entries, which were mostly Dutch words from the Spoken Dutch Corpus (Oostdijk, 2002) that were incorrectly tokenized.

In the subsequent step, we updated the lexicon to the domain of council meetings by including all words from the council transcriptions with a word frequency of 5 or higher. This resulted in the addition of approximately 15 thousand Dutch and 3,500 Frisian orthographic forms. These new forms were provided with a phonetic transcription via a grapheme-to-phoneme (G2P) conversion tool (phonetisaurus; Novak, Minematsu, & Hirose, 2015). This G2P model was trained on the FAME! lexicon and subsequently applied to the set of new words. Each orthographic form was tagged with a language label (-nl, -fr).

The final lexicon consists of approximately 190 thousand words, including 114 thousand Dutch words and 75 thousand Frisian words.

The phone set in this lexicon comprised of a complete set of 'Dutch' phones and 'Frisian' phones. This allowed for the use of specialized phones for specific Dutch and Frisian pronunciations for the AM training. The same approach was already applied in the FAME! project (Yılmaz, Van den Heuvel, & Van Leeuwen, 2018). As noted above, the orthographies were tagged with a language tag, which enabled us to differentiate between the different language routes in the ASR decoding output.

2.2.3 Language model

For the language model, a bilingual trigram model was trained with the aid of SRILM (Stolcke, 2002), with interpolated Kneser-Ney smoothing (Chen & Goodman, 1999) on the council text materials and the training set of the manual transcriptions of the council meeting recordings. Next, we created a series of interpolated LMs by mixing this new LM with the LM provided in the FAME! corpus (see Yılmaz, Van den Heuvel, & Van Leeuwen, 2018) On the basis of a grid search, we used a 0.5 weighting (thus giving equal weight to the new LM and the FAME! LM), which provided the best result (perplexity of 162.76) on a held-out test set of council text materials.

For post ASR rescoring purposes, we trained a neural net Transformer model using the Pytorch (Paszke et al., 2019) implementation that is available in the Wall Street Journal recipe in Kaldi. The Transformer model is trained on the same text materials used for RNN training in Yılmaz, Van den Heuvel, & Van Leeuwen (2018) and achieves a perplexity of 65.33 on the test set comprising transcribed council audio recordings.

2.2.4 Acoustic models

For the AM, we utilized the open source Kaldi toolkit (Povey et al., 2011) to train and test new acoustic models. We mostly followed the FAME! recipe, except we used the more recent and better performing tDNN training as provided in a Kaldi recipe for the Spoken Dutch Corpus.

We compared four different set-ups to train the recognizer, which we detail below. The different setups compare whether the new recognizer improves compared to the FAME! recognizer and subsequently tests whether adding automatically assigned speaker labels and adding additional Dutch materials improves the recognition results. Lastly, we also test whether a pure Frisian recognizer (removing any Dutch materials) influences the recognition results. Henceforth, these four set-ups will be referred to as *basic, speaker, augmented*, and *pure Frisian.*

The *basic setup* used the FAME! and council materials (Frisian and Dutch). For this setup we used the audio filename as a proxy for the speaker label in the case of the council materials, because the council materials annotations do not contain speaker labels.

In *the speaker setup* we added speaker labels to the Frisian council materials with the aid of a Pytorch convolutional neural network diarization model trained on the Spoken Dutch Corpus (see Appendix B for implementation details and performance metrics).

In *the augmented Dutch setup* we added all transcribed audio materials from the Spoken Dutch Corpus with a sampling frequency of at least 16,000 Hz to the basic setup. Lastly, in the *pure Frisian setup* we only included the audio files with the Frisian utterances from both the FAME! corpus and the Frisian council meeting training materials. In this monolingual setup, we test to what extent the bilingual setup adversely impacts recognition results compared to a monolingual Frisian setup.

3. Results

In Table 1 an overview of the word error rates (WER) is given for the different bilingual setups achieved on Frisian council meeting test set. The Augmented Dutch performed *very poorly* in an intermediate scoring step. We therefore did not complete the training for this setup and have no WER to report. The Speaker setup clearly outperformed both FAME! and the Basic setup; this recognizer was therefore used for further in-depth analysis.

Model	WER
Fame!	37.81
Basic setup	32.16
Speaker setup	**29.59**
Augmented Dutch	-

Table 1: Overview of WER for different setups on the test set of Frisian council meetings. Best performance is printed in bold.

The results presented in Table 1, are based on scoring without removing language tags from the words. Since there is an overlap between the Dutch and Frisian lexicons in terms of shared orthographic forms, this could inflate the WER purely on the basis of mismatching language tags. To study the impact of this we report the WER scores without language tags in Table 2, where we also compare the performance per language, whereby Frisian and Dutch refers to the sets of Frisian or Dutch utterances, Mix refers to the set of utterances with code-switching (i.e. both Frisian and Dutch occurs in the same utterances) and All refers to the set of all utterances.

Test set	Language	Council recognizer	FAME! recognizer
Council	Frisian	**32.05**	32.84
Council	Dutch	**19.82**	26.82
Council	mix	**36.87**	40.49
Council	all	**27.38**	31.67
FAME!	Frisian	26.39	**21.25**
FAME!	Dutch	30.30	**19.60**
FAME!	mix	36.29	**27.57**
FAME!	all	28.22	**21.65**

Table 2: WER comparison between FAME! and Council (Speaker setup) recognizer, per language for the Frisian council meeting and FAME! test sets. Language tags are ignored. Best performance is printed in bold.

Table 2 shows the WER for both the council meeting and FAME! test set. The Frisian council meeting recognizer (speaker setup) clearly outperforms the FAME! recognizer on the council test set. Conversely, the FAME! recognizer outperforms the council recognizer (speaker setup) on the FAME! test set. Lastly, the pure Frisian recognizer performs worse with a WER of 33.70 than both FAME! and the council recognizer (speaker setup) on the Frisian part of the council meeting test set.

3.1 Error Analysis

We performed a detailed analysis of the errors produced by the council recognizer (speaker setup). Word errors can be subdivided in three categories; words missing from the transcription (deletions), words which do not occur in the correct transcription (insertions), or incorrectly recognized words (substitutions). When broken down, we see that 16.27% of the reported word errors for the Frisian council meeting recognizer (all) are instances of substitutions, whereas deletions and insertions only contribute 7.49% and 3.62% respectively (summing up to 27.38%).

In our detailed error analysis, we found that the majority of ASR errors to be 'true' errors in the sense that an incorrect word was produced by the recognizer. However, there are a number of errors that are superficial in the sense that the difference between the correct word and the recognized word is so small that they should not necessarily be considered an error. We distinguish three categories of these errors: Compound words, Spelling variations, Abbreviations.

In the following sections, we describe these categories and demonstrate their impact on the WER by computing their shares on the total error rate for the council meeting test set (all).

3.1.1 Compound Words

Dutch and Frisian have the orthographic convention to join compounds into single words. For example, *crisisbeheersing* 'crisis management' consists of the stems *crisis* 'crisis' and *beheers* 'manage'. Compounding errors (e.g. transcribing *crisisbeheersing* as *crisis beheersing*) therefore have a substantial impact on the WER, because the above example results in three errors (one deletion and two insertions) for one compound word by erroneously adding a space.

This problem is particularly relevant for the domain of council meetings, which is characterized not only by frequent use of compounds but domain compound words. This increases the chance that constituents of a compound are known to the speech recognizer, but not the compound word itself. Consequently, we found that a number of compound words were recognized as multiple words (split compounds). Table 3 shows some examples of problematic compound words.

Compound	Word 1	Word 2
energiestrategie	energie	strategie
fijnstof	fijn	stof
wooncomplex	woon	complex
wijkplatform	wijk	platform
ambtenaarûndersteuning	ambtenaar	ûndersteuning

Table 3: Examples of Dutch (first four examples) and a Dutch-Frisian mixed compound (last example) where constituent words were recognized.

3.1.2 Spelling Variations

Typically, problems with spelling occur in the case of non-standard words, but it can also be the result of spelling and typing errors made by human transcribers. We identified three main sources for the spelling variation seen between the reference text and the speech recognition output. Examples of spelling alternatives encountered in the test set are presented in Tables 4 - 6.

The main cause for spelling variation errors we found is the partial overlap between the Frisian and Dutch vocabulary with minimal (or no) differences in pronunciation, but adhering to different spelling conventions (see Table 4). This means that during speech recognition the Frisian and the Dutch form of a word can become competing candidates for selection, which can result in spurious errors, i.e. if the speech recognizer chooses the Dutch form of the word, where the reference text contains the Frisian form of the word, or vice versa.

Frisian	Dutch
aginda	agenda
effekten	effecten
fan	van
inisjatyffoarstel	initiatiefvoorstel
provinsje	provincie

Table 4: Examples of Frisian and Dutch words with similar pronunciation but different orthographic representation.

Less frequently, we found a number of alternative spellings for Frisian words. An official standard spelling was introduced for the first time in 1980. In 2015 it was slightly modified (Hoekstra & Van de Velde to appear) and this reform was heavily contested by part of the language activists and Frisian writers. It should also be noted that most speakers of Frisian never write Frisian, and that on average writing skills are low (Klinkenberg et al. 2018). Only in recent years there is an increase in self-reported writing skills, due to an increase in informal writing on social media (Jongbloed-Faber 2021). Furthermore, almost all Frisians become literate in Dutch through the educational system, and the number of high school students taking Frisian as an exam subject is very low. In absence of a strong tradition of a single written standard, it is not surprising that spelling variations have perpetuated in the Frisian administrative texts used for the training of our ASR system.

Spelling Variation 1	Spelling Variation 2	Canonical Spelling
anslute	oanslute	oanslute
beantwurde	beäntwurde	beäntwurde
dankewol	tankewol	tankjewol
ferduorsaming	ferduorsuming	(neologism)
ynstânsje	instânsje	ynstânsje

Table 5: Examples of spelling variations in Frisian.

Lastly, contraction by shortening and combining words, through the omission of letters results in spelling variation. The omitted letters in the written form of the contraction are often represented by an apostrophe. For example, *zo'n* 'such a' is a contraction of the words *zo* 'such' and *een* 'a'. Contractions are more prominent in rapid speech as more sounds are reduced by the speaker. In a few cases, the contracted version of a phrase was provided by one of the transcriptions (human or ASR) while the other provided the full form.

Full Form	Contraction
yn de	yn'e
foardat	foar't
daarvoor	d'rvoor
zo een	zo'n

Table 6: Examples of contractions in Frisian (first two examples) and Dutch (last two examples).

3.1.3 Abbreviations

During council meeting abbreviations are frequently used to reference organizations and institutions. These abbreviations can be pronounced as a word, e.g. KING [kiŋ] (*Kwaliteitsinstituut Nederlandse Gemeenten* 'Institute of Quality for Dutch Municipalities), or as the sequence of individual letters e.g. BZK [bezɛtka] (*Ministerie van Binnenlandse Zaken en Koninkrijksrelaties* 'Ministry of Domestic and Royal Affairs).

When an abbreviation is known to the speech recognizer it will output a single word. If this is not the case, it is still possible for the speech recognizer to recognize individually pronounced letters which will be transcribed as a sequence of standalone letters. This creates the possibility for a discrepancy between the human transcription and recognition output. When this occurs, the ASR output is penalized on multiple accounts similar to the previously discussed compound errors.

3.2 Impact on WER

The impact of these three categories on the WER is presented in Table 7. Only those cases within the aforementioned categories that are considered correct words when the surface spelling differences are considered, have been included. The errors are distributed fairly equally between the three categories.

Category	WER	Insertion	Deletion	Subst.
Compounds	1.17	0.32	0.15	0.69
Spelling	1.96	0.01	0.02	1.94
Abbreviations	1.74	0.01	0.19	1.54
Total	4.87	0.34	0.35	4.18

Table 7: The influence of the three error categories in terms of absolute WER contributions. Subst. stands for substitution

The results reported in Table 7 indicate that the speech recognizer is more likely to produce split compounds than the human transcribers (i.e. more insertion than deletion errors). Furthermore, the ASR system produced abbreviations more often as a single word compared to the reference transcriptions (i.e. more deletion than insertion errors).

Even though the word substitutions covered by the three categories are flagged as errors during the evaluation process, it can be argued that if these substitutions would be presented to a human reader along with the corresponding audio, the reader would still consider the transcription to be correct and understand the meaning conveyed in the utterance.

When all three categories are combined they account for 4.87% of the WER (absolute). This implies that the WER would be reduced from 27.38% to 22.51% if these errors were excluded from the scoring process. As a result, this is the gain that could be achieved by addressing the orthographic differences during a post-processing step on the ASR output.

The remaining WER (22.51%) represents the percentage of true errors being produced by the speech recognition system. While these errors cannot easily be addressed, an analysis of them still provides useful insight into the weaknesses of the speech recognition system.

We have identified that word forms distinguished by a suffix present a challenge for the speech recognizer to recognize correctly, more particularly in cases where suffixes represent morphemes with tense information (verbs) or plurality (nouns), where the word class remains the same and the competition cannot be resolved by the language model. In casual or rapid speech, the ending of words is often reduced, meaning they are not pronounced fully or are even dropped to create a fluid connection with the following word. In Table 8 we have listed some of the word forms which were substituted in the recognition of the test set, which account for another 0.83% of the WER (absolute).

Reference	**Hypothesis**
besluit	besluiten
denk	denken
fierder	fierders
ferantwurdlik	ferantwurdlike
herinner	herinnerje
moat	moatte

Table 8: Examples of suffixation errors in Dutch (first two examples) and Frisian (last four examples).

4. Discussion and Conclusion

The aim of this project was to create an ASR recognition system for Frisian council meeting speech. The new ASR system for Frisian must be able to automatically recognize audio recordings of the many Frisian councils and handle the Frisian/Dutch code-switching typical in these settings. To that end, the original FAME! ASR system was taken as a starting point. Four different extensions were tested, primarily varying in the way how additional audio and text materials were added. For the best extension, the recognition errors were analyzed in terms of insertion, deletion and substitution errors and hypothetical values if the lexicon were ideally cleaned up in terms of compounds, spelling variations and abbreviations.

Within the project, a corpus of domain specific texts from the Frisian councils was collected. We tried to collect as much Frisian language materials as possible. About 10% of the domain specific materials was in Frisian, 90% in Dutch, which reflects the dominance of Dutch in the local and regional administration. Thanks to a special transcription platform and a training program for transcribers we were able to quadruple the amount of manually transcribed Frisian speech compared to the materials available in the FAME! corpus.

The results show that adding more (in domain) audio and text material helps to improve the system, and that this must be done in a controlled and balanced manner in order to take effect. For reducing the word error rate, balancing is more relevant than just more data. In addition, error analyses showed that adding compounds and abbreviations to the lexicon will further reduce the word error rate with a few absolute percentage points. Dealing with spelling differences between Frisian and Dutch will also help to reduce the word error rates.

Practical options for further improvement of the ASR system as a whole include the improvement of the AM by using modern neural network approaches. The phone sets of Frisian and Dutch are currently completely separate as are the lexica. Merging these sets in an intelligent way (e.g. using the same consonant phones for Frisian and Dutch) might further improve decoding results. In addition, for bilingual meetings it is worthwhile to consider a genuinely bilingual ASR system in which Dutch and Frisian LMs are put in parallel. The use of more audio and text data will make sense as long as this happens in a balanced way. Finally, the AM might profit from a cleaning-up of the phone symbols used in the lexicon from the FAME! system that we started with.

The audio recordings and their transcriptions will be made available as the Frisian Council Meetings Corpus (FCMC) via CLARIN Data Centre INT[1].

The present version of the ASR will be used to provide initial transcriptions in a subtitling service for the audio/video stream of the council meetings which will be uploaded to a cloud service. In this way, Frisian council meetings will conform to the legal obligation to make recordings of public council meetings accessible for the deaf and hard of hearing.

5. Bibliographical References

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359-393.

Hoekstra, E. & H. Van de Velde (to appear). The History and Standardization of Frisian. In: *From Arantzazu to the the World (1968-2018).* Bilbao, Euskaltzaindia.

Jongbloed-Faber, L. (2021). Frisian on social media. The vitality of minority languages in a multilingual online world. LOT, Amsterdam.

Novak, J.R., Minematsu, N., & Hirose K. (2015). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, pp. 1–32.

[1] https://centres.clarin.eu/centre/22

Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus In: Peters, P., Collins, P., Smith, A. (ed.), *New frontiers of corpus research*. Papers from the twenty first international conference on English language research on computerized corpora, Sydney 2000, pp. 105-112.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*, 8026-8037.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Yilmaz, E., Derinel, A., Kun, Z., Van den Heuvel, H., Brummer, N., Li, H., & David van Leeuwen (2019). Large-Scale Speaker Diarization of Radio Broadcast Archives. *Proceedings Interspeech 2019*, Graz, Austria, 15-19 September 2019.

Yilmaz, E., Dijkstra, J., Van de Velde, H., Kampstra, F., Algra, J., Van den Heuvel, H. & Van Leeuwen, D.A. (2017). Longitudinal Speaker Clustering and Verification Corpus with Code-switching Frisian-Dutch Speech. *Proceedings of Interspeech 2017,* pp. 37-41.

Yilmaz, E., McLaren, M., Van den Heuvel, H., & Van Leeuwen, D.A. (2017). Language Diarization for Semi-Supervised Bilingual Acoustic Model Training. *Proceedings of ASRU 2017. IEEE Automatic Speech Recognition and Understanding*. IEEE.

Yılmaz, E., McLaren, M., Van den Heuvel, H., & Van Leeuwen, D. (2018). Semi-supervised acoustic model training for speech with code-switching, *Speech Communication*, Volume 105, December 2018, pp. 12-22. DOI: https://doi.org/10.1016/j.specom.2018.10.006

Yilmaz, E., Van den Heuvel, H., & Van Leeuwen, D.A. (2017). Exploiting Untranscribed Broadcast Data for Improved Code-switching Detection. *Proceedings of Interspeech 2017*, pp. 42-46.

Yılmaz, E., Van den Heuvel, H., & Van Leeuwen, D. (2018). Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech. *Proceedings Interspeech 2018*, September, 2-6, 2018, Hyderabad, India.DOI: 10.21437/Interspeech.2018-52

6. Language Resource References

Fryske Akademy (2021). Frisian Council Meetings Corpus. Distributed via Clarin Data Centre INT, https://centres.clarin.eu/centre/22

7. Appendix A: Language Classification

We trained a language classifier with the scikit learn toolkit, using Naive Bayes. The implementation was inspired by the following web page: https://towardsdatascience.com/an-efficient-language-detection-model-using-naive-bayes-85d02b51cfbd, and the final implementation details can be found here: https://github.com/martijnbentum/frisian_asr/blob/main/utils/language_detection.py

We trained and utilized multiple classifiers to label the text materials. The language classifier becomes more accurate when it receives more data to classify. A whole text would be easier than a sentence, which in turn would be easier than a single word. However, with more material (e.g. a whole text or sentence), you lose specificity. All words in a text or sentence are classified as a specific language. Since code switching with Dutch occurs frequently in Frisian language use we need word level specificity. To balance these conflicting conditions, we trained both sentence and word level classifiers. We trained the classifiers on two thirds of the data and tested them on the held-out test set.

language	precision	recall	F1
Dutch	0.96	0.96	0.95
Frisian	0.97	0.94	0.95

Table 1: Performance results of the sentence level language classifier.

language	precision	recall	F1
Dutch	0.83	0.78	0.81
Frisian	0.80	0.84	0.82

Table 2: Performance results of the word level language classifier.

language	precision	recall	F1
Dutch	0.96	0.95	0.95
Frisian	0.97	0.97	0.97

Table 3: Performance results of the word level language classifier, omitting words that occur both in Frisian and Dutch.

language	precision	recall	F1
Dutch	0.96	0.96	0.96
Frisian	0.97	0.97	0.97

Table 4: Performance results of the combined word and sentence level language classifier.

Table 1 - 3 show the classification results on the held-out test set. The difference between sentence level and word level classification can be seen in Table 1 and 2. However, it is important to realize that some words occur both in Frisian and Dutch. It is difficult for the classifier to distinguish between these words at the word level. Table 3 shows that the performance improves when only testing words that occur either in Dutch or Frisian: A substantial gain in F1 can be observed. To alleviate the drop in performance for the word level classifier we combined the word and sentence level classifier.

The combined classifier uses a two-step classification process. In a first classification pass, sentences are classified. Subsequently, for each word not occurring in both Frisian and Dutch, the word level classifier classifies the word, all other words are classified as the language of the sentence. The overall results of this combined classifier outperforms the other classifiers (see Table 4)

8. Appendix B: Speaker Labelling

The Frisian council meeting transcriptions do not contain speaker labels. To investigate whether we can improve recognition results by automatically labelling speakers we trained a convolutional neural network (CNN) model inspired by this github repository: https://github.com/WiraDKP/pytorch_speaker_embedding _for_diarization/tree/master/src, the final implementation details can be found here: https://github.com/martijnbentum/frisian_asr/tree/main/SPEAK_RECOG.

We used all recordings in the Spoken Dutch Corpus that are longer than 2 seconds and have a sample rate of 16000 Hz or higher. The remaining recordings include approximately 3,000 distinct speakers.

# Speakers	precision	recall	F1
1	0.92	0.88	0.90
2	0.79	0.83	0.81
3	0.66	0.75	0.70
4	0.46	0.77	0.58
5	0.36	0.49	0.41
6	0.23	0.70	0.34
7	0.38	0.32	0.35
8	0.09	0.26	0.14

Table 1: Overview of precision and recall of speaker labelling of Spoken Dutch recordings

The Spoken Dutch corpus contains speech recordings with a different number of speakers in them. Table 1 gives an overview of the performance. The scores for one-person files is not perfect because the model assigned multiple labels to this single speaker. The overall accuracy is 86%.

The speaker labelling improved the performance results of the Frisian council recognizer. However, this was dependent on the length of the audio stretches that were used. When labelling speakers for an entire council meeting, 1 to 4 hours of recordings, the speaker labelling did not improve recognition results. It was only when using intermediate audio files of 1 - 10 minutes (most audio files are about a minute), already segmented out of the longer meeting recordings, that the automatic speaker labelling improved the recognition results.

9. Acknowledgements

The development of the speech recognizer was financed and executed by Provinsje Fryslân, Wetterskip Fryslân and the municipalities Achtkarspelen, Dantumadiel, Fryske Marren, Heerenveen, Leeuwarden, Noardeast-Fryslân, Opsterland, Smallingerland, Sudwest-Fryslân, Tytsjerksteradiel and Waadhoeke.

The development of the multilingual and multidialect transcription software was financed and executed by Fryske Akademy and Humain'r.

The development of the training program for Frisian/Dutch transcriptions for training ASR systems was financed and executed by Fryske Akademy and Humain'r.