The Boarnsterhim Corpus: A Bilingual Frisian-Dutch Panel and Trend Study

Marjoleine Sloos, Eduard Drenth, Wilbert Heeringa
Fryske Akademy, Royal Netherlands Academy of Sciences (KNAW)
Doelestrjitte 8, 8911 DX Ljouwert, The Netherlands
{msloos, edrenth, wheeringa}@fryske-akademy.nl

Abstract
The Boarnsterhim Corpus consists of 250 hours of speech in both West Frisian and Dutch by the same sample of bilingual speakers. The corpus contains original recordings from 1982-1984 and a replication study recorded 35 years later. The data collection spans speech of four generations, and combines panel and trend data. This paper describes the Boarnsterhim Corpus halfway the project which started in 2016 and describes the way it was collected, the annotations, potential use, and the envisaged tools and end-user web application.

Keywords: West Frisian, Dutch, sociolinguistics, language variation and change, bilingualism, phonetics, phonology

1. Background

West Frisian is mostly spoken in the province of Fryslân in the north of the Netherlands. All its speakers are bilingual with Dutch, which is the dominant language. West Frisian is mainly used in informal settings (van Bezooijen 2009:302) but also the most formal, viz. in the provincial parliament. In semi-formal interactions (e.g.shopping, in church), Dutch is usually the preferred language. This dominant position of Dutch traces back to about 1500 when Fryslân lost its political independence and the upper class started to use a mixed Frisian-Dutch language (van Bezooijen 2009:302). Given this long-term contact, it was (and still is) often thought that West Frisian is slowly but steadily converging towards Dutch (e.g. Feitsma 1989).

However, an investigation in the 1980s into phonological change in West Frisian and Dutch of (the same) West Frisian speakers, suggests that, at least at the phonological level, the opposite holds: younger speakers were more likely to keep the phonological rules of the two languages apart; whereas older speakers were more likely to confuse them in either language (van der Kuip 1986, Feitsma et al. 1987, Feitsma 1989, Meekma 1989). This suggests language change, and we are currently investigating whether this change has been continuing over the past thirty-five years in a follow-up study—which is a replication of the original one (see also section 2).

The present contribution describes the unique dataset that underlies these studies. The database is longitudinal, spanning four generations. It combines trend data with a panel study in which the same speakers were recorded in the 1980s and 35 years later. For this sociolinguistic study, unique in a bilingual context, data were collected that have never been made publicly accessible before. Given the special language situation, the exclusive design, and the broad array of linguistic fields which could benefit from these data, we will make the corpus freely accessible.

The remainder of this paper describes the data collection and methodology in section 2. Section 3 describes the embedding in a larger infrastructure and section 4 provides background information about the tool that is used to retrieve lexical frequency. Section 5 discusses previous and ongoing research, and further research opportunities that this database may make possible. Finally, section 6 concludes.

2. Methods and Data

2.1 Data Collection

The Boarnsterhim Corpus (henceforth BHC) consists of two collections. BHC1 was recorded between 1982 and 1984; BHC2 is recorded in 2018-2019. The BHC1 data underlie the above mentioned sociolinguistic studies into variation and change of bilingual Frisian-Dutch speakers (van der Kuip 1986, Feitsma et al. 1987, Feitsma 1989, Meekma 1989). The recordings were made in the municipality of Boarnsterhim in central Fryslân (see Figure 1), because the inhabitants—more than other Frisians—advocated the monolinguistic use of Frisian (Feitsma 1989).[1]

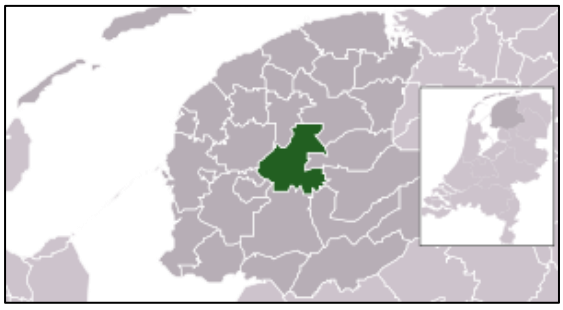

Figure 1: The municipality of Boarnsterhim 1984-2014. The inset represent the Netherlands (*Centraal Bureau voor de Statistiek* 'Statistics Netherlands').

[1] The municipality of Boarnsterhim (Dutch: Boornsterhem) was created through a division boundary alteration in 1984 in which three municipalities were combined into one. The area was 168.58 km^2 and consisted of 18 villages of which Grou was the principal one. It is an area with much water (17.04 km^2) and remained relatively isolated for quite a long time. In 1990, the population consisted of 17,710 people and slightly increased every year. As a result of another division boundary alteration in 2014, Boarnsterhim was again divided and combined with four other municipalities.
(https://nl.wikipedia.org/wiki/Boornsterhem.)

The speakers were recorded twice; once in West Frisian (with a West Frisian native interviewer) and once in Dutch (with a monolingual Dutch interviewer). To enhance informal and spontaneous speech—crucial for Frisian—the recordings took place at the speakers' homes. This lead to a variable amount of noise in the recordings, but overall the quality of the recordings is still acceptable and even valuable for phonetic research.

Within the same family, three speakers were recorded, each of a different generation. The three speakers of a single family were either male or female. The older speakers in BHC1 were born between 1898 and 1917 (65-86 years old at the time of the recording); middle-aged speakers were born between 1926 and 1946 (36-58 years old at the time of the recording); and younger speakers were born between 1952 and 1962 (20-32 years old at the time of the recording) (Feitsma et al. 1987, Feitsma 1989, Meekma 1989). In sum 87 speakers were recorded, from 29 families (two speakers were missing).

The recorded family members had comparable socio-economic status (SES) and were socially divided into three categories based on their levels of education: non-educated farmers, lower educated, or higher educated. The non-educated females were wives and daughter of farmers. Higher educated females of the older generations appeared too difficult to find at the time, so the social stratification in the BHC1 corpus consists of five groups (van der Kuip 1986, Feitsma et al. 1987, Feitsma 1989, Meekma 1989):

- higher-educated males
- lower-educated males
- lower-educated females
- male farmers (non-educated)
- non-educated females

Each recording consists of 20 read sentences, a read story (2-3 minutes), and an interview of about 40 minutes about the speaker's use of West Frisian, language attitude, and daily life activities. The data were originally recorded on TDK-AD (Japan) cassette tapes and digitalized in 2016 with a SONY TC-FX310 cassette deck connected to a PC with a stereo Jack-Tulp cable.

The BHC2 aims to be a replication of the BHC1, with the same number of speakers and same age groups. Since education levels gradually increased and all farmers are (lower or higher) educated nowadays, we adhere to the currently common two-way distinction in education level (and SES in general) in the Netherlands.[2]

The generation of the middle-aged speakers of the BHC1 serves as the oldest generation in the BHC2 and the generation of the youngest speakers of the BHC1 can be identified with the middle-aged speakers in the BHC2. Our aim is to record 50% of the original speakers of the BHC1 and 50% new speakers of these generations, plus an equal number of younger speakers born between 1982 and 1997. We follow the system of three family members of the same gender throughout the data collection. The recordings are made with Tascam DR-44WL recorders.

2.2 Annotation and Data Labelling

The data are manually aligned and annotated in Praat speech processing software (Boersma & Weenink 2017). The textgrids consist of orthographic, lexical, phonological, and phonetic annotations. The orthographic transcriptions are aligned at the phrase level in Standard West Frisian (cf. the *foarkarswurdlist* 'preferred wordlist' 2011) or Standard Dutch (cf. *Het Groene Boekje* 'The Green Booklet' 2017). Dutchisms in Frisian recordings and Frisisms in Dutch recordings are given between square brackets.

Separate tiers are made for the alignment and annotation of words, phonemes, and phonetic or allophonic realizations. A point tier is used to indicate deletion. Extra tiers for specific phonological processes will be added in the future, like a tier for the pronunciation of final -ən. In addition, a tier is used for general comments (see Figure 2).

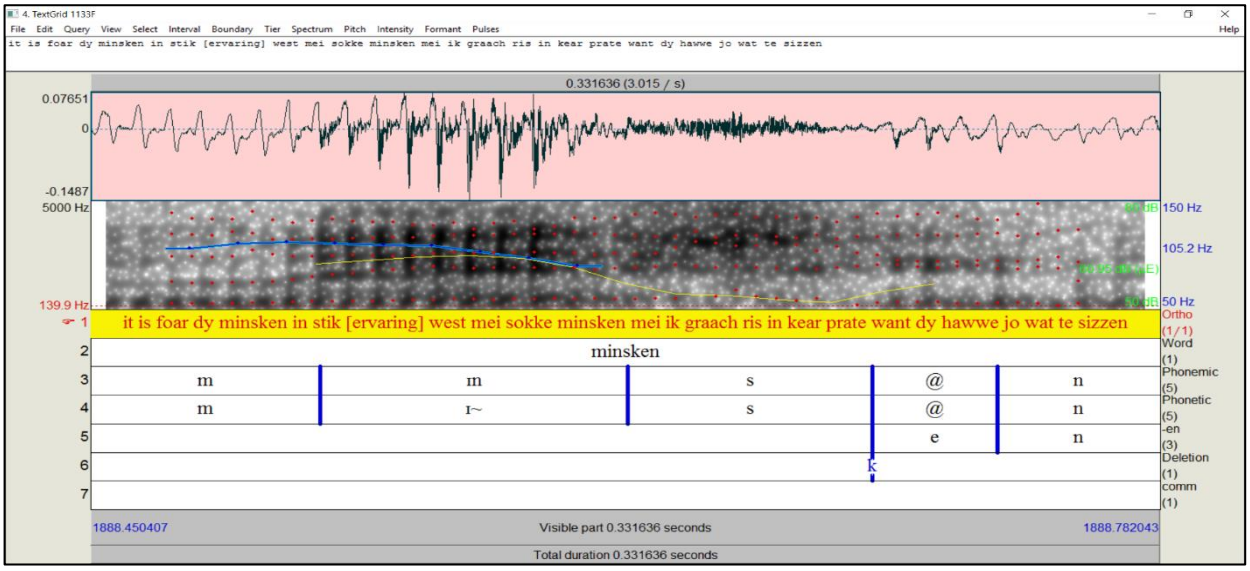

Figure 2: A textgrid from the BHC1.

[2] In the BHC1, higher education correspond to ULO (extended elementary education) for older speakers, and HAVO (senior general secondary education). Since the general education level further increased over the past decades, the cut-off point for the BHC2 is higher, i.e. HBO (higher vocational education) and WO (university) are regarded higher education.

3. Infrastructure

The BHC (1 &2) will be embedded in a large-scale Clarin infrastructure in the Netherlands (Odijk & van Hessen 2017), as part of the European digital infrastructure for the humanities CLARIAH (Odijk 2016). This provides the opportunity to connect information from the BHC to frequency and POS tagging of other databases as well. The corpus will be encoded in Text Encoding Initiative (TEI) (Ide & Véronis 1995, Sperberg-McQueen & Burnard 1995). The different parts of this portal will be gradually published online in the near future.

The BHC will be made available in different phases. Initially, the orthography (and other information in the textgrids) will be published and POS-tagged, so that it can be queried. Subsequently, in later stages, original recordings and linguistic metadata (anonymized speaker data), will be made available. Finally, textgrids and an XML version of them will be added. A webpage with general background information about the corpus is made available at the website of the Fryske Akademy.[3]

4. Frequency Analysis Tool

An important factor in language variation and change is lexical frequency (e.g. Bybee 2006, Phillips 2006, Diessel 2007). Therefore, it is crucial to offer frequency information to the end user. [4] But the corpus size of the complete BHC is nearly 125 hours of speech for both West Frisian and Dutch. Frequency based solely on the BHC may be biased, also because the population sample in the BHC is deliberately homogeneous, and the variety of topics discussed in the recordings restricted. To obtain more reliable frequency data, we turn to The Frisian Audio Mining Enterprise (FAME!) (Yilmaz et al. 2016). FAME! is a corpus consisting of more than 2600 hours of West Frisian Radio Broadcast from 1950-2016. The Corpus Spoken West Frisian contains orthographical annotations of speech by 60 speakers, recorded between 2003 and 2006. We will provide both token and lemma frequency of words.

The BHC will be both lemmatized and POS-tagged by using TreeTagger (Schmid 1994, 1995) or Frog (Avontuur et al. 2012). POS-tagging of the data makes it possible to easily find words of a particular word class, for example verbs, in the corpus. No lemmatizer/POS-tagger for West Frisian exists, but given that the morphological and syntactic structure of West Frisian and Dutch are highly similar, the West Frisian speech data will be translated from West Frisian to Dutch before being POS-tagged. We will implement an application which automatically translates Frisian to Dutch by using the existing API of the online translation system Oersetter (lit. 'Translator').[5]

For the Dutch part, we refer to frequency information in the existing Dutch corpora: the Corpus Spoken Dutch (*Corpus Gesproken Nederlands*) (Oostdijk 2000, 2003, Oostdijk & Broeder 2003), CELEX (Baayen et al. 1993), and Nederlab (Brugman et al. 2016). These corpora also provide token frequencies and lemma frequencies.

5. Research based on the BHC

5.1 Previous research

The BHC1 corpus was partly analysed (35 males and 8 females in separate studies) for five phonological rules in West Frisian:

- schwa deletion in final -/ən/
- nasal place assimilation in final -/ən/
- vowel nasalization
- assimilation of final -/s/ to [z d Ø]
- initial /d/ deletion of the definite non-neuter article [də]

Feitsma et al. (1987) and Feitsma (1989) report that among the male speakers, assimilation of -s and d-deletion occurred more in Frisian than in Dutch. They produced as many nasalized vowels in Dutch as in West Frisian (whereas in Standard Dutch nasalization does not occur). Younger speakers nasalized even more than older speakers, especially in Dutch, pointing at a transfer from Frisian to Dutch. Another age difference was observed for schwa-deletion in final -/ən/: younger speakers used more schwa deletion (the West Frisian rule) in West Frisian and n-deletion (the Dutch rule) in Dutch, whereas the older speakers mixed up the rules in the two languages more often. In this case, the younger speakers kept the two rules better apart than the older speakers. A later study on schwa deletion of eight females pointed towards the same direction (Meekma 1989).

5.2 Current research

Currently, we are investigating the large array of pronunciations of final -/ən/ based on the BHC1 in Frisian and Dutch in more detail. We also investigate the interaction between the final nasal that remains after ə-deletion and the preceding consonant(s). We study the phonological status of the remaining nasal after ə-deletion with regard to syllabicity (Sloos et al. submitted).

5.3 Future research possibilities

The BHC offers a wide range of applications for linguistic research. We plan a sociophonetic study into variation and change in final -/ən/. This will later be extended to a broader investigation into the variation and change of the phonological system of Frisian and Dutch of Frisian-Dutch bilinguals across four generations, spanning the pronunciation of speakers born between 1897 and 1997. The BHC also offers data for a comparative study, based on the two varieties, giving more insight into convergence and divergence in a bilingual contact situation in which one language is dominant.

Secondly, current phonological descriptions of West Frisian lack a phonetic basis. The BHC contains ample material for phonological (re)investigation.

Thirdly, since the same speakers are investigated in the two languages, this corpus serves studies into bilingualism. In addition, the West Frisian-accented Dutch data could be compared to other varieties Dutch, for instance, in the Spoken Dutch Corpus (CGN).

[3] https://www.fryske-akademy.nl/.

[4] The BHC1 studies did not take frequency into account as a factor of variation.

[5] http://oersetter.nl/ accessed at September 2017.

Fourthly, this corpus provides the opportunity to investigate real-time language change, and compare real-time and apparent-time studies into language change.

Finally, the corpus may facilitate studies into language attitude—and changing attitudes—toward the usage of Frisian by first- and second language users. Also the development of reading skills in West Frisian could be investigated, which is interesting given the increasing efforts regarding bilingual education at all levels in the province of Fryslân.

6. Conclusion

We have described in detail our efforts to make available a sociolinguistic corpus of West Frisian and Dutch of the same bilingual West Frisian-Dutch speakers. The corpus contains data for four generations (born between 1897 and 1997). One part consists of speech of speakers that were recorded twice, with an interval of more than 30 years. Along with the sound files and textgrids with much phonological information, anonymized metalinguistic information of the speakers will be made available for research purposes. We will also provide lexical frequency information (token and lemma frequency).

7. Acknowledgements

This research has been made possible through a VENI grant (number 275-75-10) by the Netherlands Organization for Scientific Research to the first author, matched by the Fryske Akademy, which is gratefully acknowledged. The BHC1 studies were funded by *Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek* (currently Netherlands Organization for Scientific Research), Stichting Taalwetenschap Fryske Akademy, and the Fryslân Bank.

8. Bibliographical References

Avontuur, T., Balemans, I., Elshof, L., van Noord, N., & van Zaanen, M. (2012). Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal* 2: 34—51.

Baayen, R., Piepenbrock, R. & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Boersma, P. & Weenink, D. (2017). Praat, doing phonetics by computer.Version 6.0.30. www.praat.org.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E. & van den Bosch, A. (2016). Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. *LREC Proc. 2016.* 1277—1281.

Bybee, J. (2006). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25(2): 108—127.

Feitsma, A. (1989). Changes in the pronunciation of Frisian under the influence of Netherlandic. In Deprez, K. (ed.), *Language and Intergroup Relations in Flanders and in the Netherlands*, 181—193. Dordrecht: Foris.

Feitsma, T., van der Geest, E., van der Kuip, F.J. & Meekma, I. (1987). Variations and development in West Frisian sandhi phenomena. *International Journal of the Sociology of Language* 64: 81—94.

Ide, N., & Véronis, J. (Eds.). (1995). *Text encoding initiative: Background and contexts* (Vol. 29). Springer Science & Business Media.

Meekma, I. (1989). Frouljuspraat en it lytse ferskil. Oer útspraakferoaring yn 'e sandhi by froulju en manlju. *It Beaken* 51: 115—129.

Odijk, J. (2016). CLARIAH in the Netherlands. *Proceedings of LREC 2016*.

Odijk J. & van Hessen A. (2017). CLARIN in the Low Countries. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi

Oostdijk, N. (2000). The Spoken Dutch Corpus Project. In *The ELRA Newsletter*. Vol. 5 No. 2: 4—8.

Oostdijk, N. (2003). Het Corpus Gesproken Nederlands: Veelzijdig onderzoeksinstrument voor o.a. taalkundig en taal-en spraaktechnologisch onderzoek. In *LINK* 14(1): 3—6.

Oostdijk, N. & D. Broeder (2003). Een databank van het gesproken Nederlands. In *Philologia Frisica 2002*: 37—54.

Phillips, B. (2006). *Word frequency and lexical diffusion*. New York: Palgrave Macmillan.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Sloos, M., Ariza Garcia, A. & van de Weijer, J. (submitted). Syllabic vs. non-syllabic nasals in West-Frisian. *Journal of the International Phonetic Association.*

Sperberg-McQueen, C. M., & Burnard, L. (1995). The design of the TEI encoding scheme. *Computers and the Humanities*, 29(1), 17—39.

van Bezooijen, R. (2009). The pronunciation of /r/ in West Frisian. In J.N. Stanford & D.R. Preston (Eds.), *Variation in Indigenous Minority Languages* (Studies in Language and Society 25). Amsterdam and Philadelphia: John Benjamins, pp. 299—318.

van der Kuip, F. J. (1986). Syllabisearring yn it Frysk en it Hollânsk fan Fryskpraters. *Tydskrift foar Fryske Taalkunde* 2: 69—92.

Yilmaz, E., Andringa, M., Kingma, S., Dijkstra, J., van der Kuip, F., Van de Velde, H. & van Leeuwen, D. (2016). A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research. *LREC Proc. 2016.* 4666—4669.