Developing Infrastructure for Low-Resource Language Corpus Building

Hedwig Sekeres[1], Wilbert Heeringa[1,2], Wietse de Vries[1], Oscar Yde Zwagers[1], Martijn Wieling[1], Goffe Th. Jensma[1]
University of Groningen[1], Fryske Akademy[2]
Broerstraat 5, 9712 CP Groningen[1], Doelestraat 8, 8911 DX Leeuwarden[2]
{h.g.sekeres, wietse.de.vries, o.y.zwagers, m.b.wieling, g.t.jensma}@rug.nl,
wheeringa@fryske-akademy.nl

Abstract

For many of the world's small languages, few resources are available. In this project, a written online accessible corpus was created for the minority language variant Gronings, which serves both researchers interested in language change and variation and a general audience of (new) speakers interested in finding real-life examples of language use. The corpus was created using a combination of volunteer work and automation, which together formed an efficient pipeline for converting printed text to Key Words in Context (KWICs), annotated with lemmas and part-of-speech tags. In the creation of the corpus, we have taken into account several of the challenges that can occur when creating resources for minority languages, such as a lack of standardisation and limited (financial) resources. As the solutions we offer are applicable to other small languages as well, each step of the corpus creation process is discussed and resources will be made available benefiting future projects on other low-resource languages.

Keywords: low-resource language, online corpus, corpus creation

1. Introduction

This paper introduces the infrastructure and software used to create a monolingual diachronic corpus for an under-resourced language variety. The corpus was created for Gronings, a language variety spoken in the north of the Netherlands, and is freely accessible as part of a larger online database on this language variant, called WoordWaark. This paper will detail the steps taken in the creation of this corpus and offer recommendations for future corpus building projects in order to also benefit other minority languages.

Gronings is a variant of the Low Saxon language, which is spoken in the Netherlands and Germany and is recognised within the Netherlands under Part II of the European Charter for Regional or Minority Languages (ECRML, 1998). Although exact numbers of speakers are difficult to determine, variants of Low Saxon in the Netherlands are in decline and show clear age-grading, with only a relatively small proportion of young speakers (Bloemhoff, 2005; Versloot, 2020). As intergenerational transmission of the language within families is declining, it is imperative that resources facilitating both research and language learning are created. As of yet, no indexed corpus for written Gronings exists. Although attempts have been made to standardise the spelling of Gronings (e.g., Ter Laan, 1947; Reker, 1984), it can hardly be considered a standardised language. These attempts take the form of a set of guidelines rather than strict rules as authors writing in Gronings often want to reflect their (local) pronunciation of a word in its spelling. Additionally, these spelling guidelines are not always known or accepted by everyone who produces writing in Gronings. Both of these factors cause a substantial amount of spelling variation, which is increased in our corpus by language change in general, which is also reflected in the spelling.

Although there have been developments in the collection of written corpora for languages without a standardised orthography (e.g., Millour and Fort, 2020), previous endeavours in creating annotated corpora for minority languages (e.g., Linder et al., 2019; Tracey et al., 2019; Tahir and Mehmood, 2021) usually do not address the challenges that internal variation poses for developing language technology, which do not only apply to Gronings but to many minority languages. Although spelling variation can pose a challenge for corpus creation, this is not to say that spelling variation in itself is negative or harmful to language preservation or emancipation. In fact, retaining regional, diachronic and idiosyncratic spelling variation as found in the original texts is one of the main features of our corpus.

The written corpus created in this project is an integral part of WoordWaark, an online openly accessible language database for Gronings which interlinks, among other things, several dictionaries, survey data on language variation, and (audio) material contributed by speakers of the language. As of January 2024, the corpus contains

10,036,643 tokens, 243,466 types and 622,470 sentences from 431 documents. As a part of WoordWaark, the corpus serves two main goals. On the one hand it facilitates linguistic research on Gronings. On the other hand it makes the body of written texts in Gronings accessible to a general audience. For the first goal, it is necessary that the corpus includes sufficient linguistic information, such as part-of-speech tags, and that it presents sentences exactly as they were found in the original texts. For the second goal, it is important that the sentences in the corpus can be used to illustrate KWIC-entries from the dictionary and thereby be used by a general audience as a reference work, to broaden their knowledge of real-life applications of words found in the dictionaries and as a tool to learn the language.[1]

In addition to serving different audiences with one corpus, the method proposed here is particularly suited to contexts of (financially) under-resourced languages as it makes use of volunteers and automation, thereby both involving the speaker community in the preservation of language, and reducing the amount of labour necessary.

2. Requirements

2.1. Texts

Several materials need to be in place or be arranged in order to build a corpus of this type. First and foremost, a collection of written texts in the target language is needed. The texts used for the WoordWaark corpus were available through the Library of the University of Groningen. All texts that were tagged with the word 'Gronings' were included in our initial search, resulting in 763 texts, containing published books, periodicals, magazines, posters and miscellaneous publications ranging in publication year from 1822 to 2016. This also meant that some texts that were erroneously tagged with Gronings but were actually a different Low Saxon dialect or texts that were about the province of Groningen but not written in Gronings had to be later excluded, and that there might have been texts that were (partly) written in Gronings that were not tagged as such that were therefore not included. All (included) texts that are still copyrighted (all but 124) are not published integrally, but only cited from their original works as KWICs and publicly searchable but not downloadable. Although for many corpora, it is important to be restrictive in the selection of texts in order to ensure that the corpus is balanced and representative of different types of texts (Ädel, 2020), this is less feasible for low-resource varieties such as Gronings, for which all available printed text need to be included in order to keep a substantial corpus. All texts were already assigned a unique identifier by the University Library, and had some metadata associated (such as title, author(s), publisher, etc.). Through the identifiers, it was possible to request texts in batches from the University Library so that volunteers could process them, and to keep track of the status of each text in the pipeline.

2.2. Volunteers

The second requirement for building the corpus is to have an organisation that is capable of recruiting and coaching volunteers. For this project, it was not necessary for all volunteers to be proficient in Gronings, but most of them were. Proficiency in Gronings was most useful when there was doubt about the dialect of Low Saxon a text was written in, but was not necessary for either adding metadata, or checking and correcting the optical character recognition (OCR) results after scanning the texts in print. A total of 13 volunteers worked on this project, although not all simultaneously. Most of the volunteers were retirees with active or passive knowledge of Gronings, who had an interest in language and literature in general. An exception were the volunteers who scanned books, as elderly volunteers were hesitant to perform in-person tasks due to the COVID-19 pandemic and student volunteers were recruited instead. Volunteers were recruited through the Center for Groningen Language and Culture as well as through the Dutch heritage platform *Erfgoedvrijwilliger*.[2] Volunteers were offered a small hourly compensation for their work, in accordance with the Dutch Tax and Customs Administration. Volunteers that did tasks from their own home (relating to OCR and metadata) were provided with a laptop where all required software was installed, which also included TeamViewer, so that help could be provided and the computer could be controlled remotely if the volunteers encountered problems or had questions. We estimate that volunteers have spent between 1800 and 2000 hours working on the corpus thus far. One member of the project team was available through email and telephone to answer questions and solve problems for the volunteers.

2.3. Digital Infrastructure

The final requirement for building this corpus was to have a digital infrastructure in place in order to ensure a smooth process combining work done by volunteers and automation. This digital infrastructure consisted of a pipeline which all texts went

[1] The corpus will also be included in a massive open online course for Gronings to provide resources to new speakers.

[2] `www.erfgoedvrijwilliger.nl`

through. Each step of this pipeline (see Figure 1) will be explained in detail below.

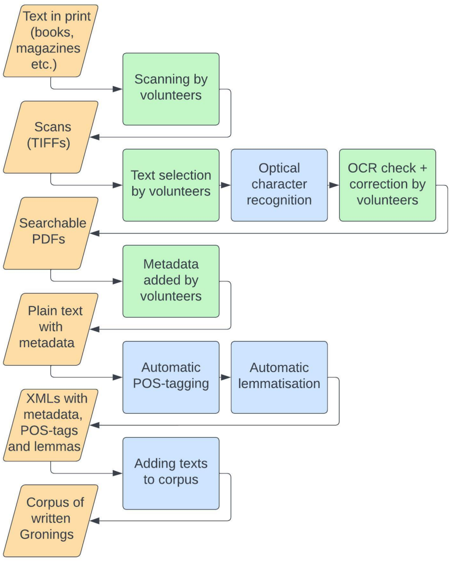
Text in print (books, magazines etc.)
Scanning by volunteers
Scans (TIFFs)
Text selection by volunteers
Optical character recognition
OCR check + correction by volunteers
Searchable PDFs
Metadata added by volunteers
Plain text with metadata
Automatic POS-tagging
Automatic lemmatisation
XMLs with metadata, POS-tags and lemmas
Adding texts to corpus
Corpus of written Gronings

Figure 1: Pipeline used for converting texts in print to a corpus. Green boxes represent steps conducted by volunteers, blue boxes represent automated steps.

3. Volunteer Tasks

3.1. Scanning

The first step in the pipeline was to create digital scans from the texts. Volunteers came to the University Library (UL) and were instructed to scan the texts from cover to cover, using a CZUR-ET16 overhead scanner. Although only running text would be used in the final corpus, the inclusion of the covers and first and last few pages of all included books helped with the retrieval of relevant metadata later in the process. The scans were saved using the unique UL identifier and exported as colour TIFF files with LZW compression and stored in a Google Drive folder. Some of the texts were difficult to scan using the overhead scanner because of issues with light reflecting from pages or books having rigid spines. These texts were scanned using a Ricoh MP C3003 multi-function (flatbed) printer, at 300 dpi, in black-and-white and at full brightness (these settings proved to deliver the best quality scans for OCR). These scans were also saved as TIFF files using the unique identifier and exported to the Google Drive folder. The quality of these scans was lower than those of the CZUR scanner, but still sufficient (using the aforementioned settings) to conduct OCR.

3.2. Text Selection and Correction

The next step in the pipeline was to convert the scans to text using optical character recognition (OCR), using ABBYY FineReader 15 Corporate. First, the volunteers indicated the text areas that needed to be converted, which meant selecting and deselecting areas so that only running text in Gronings remained. In other words, all areas that were not text (e.g., images or page numbers), that were not Gronings (e.g., parts of multilingual texts in, for example, Dutch or other Low Saxon dialects) or not running (e.g., tables, word lists, title page, chapter titles, etc.) had to be deselected as we are only interested in full sentences for this corpus. Then the volunteers had to instruct the program to start converting the selected areas to text. In advance, we provided the program with a lexicon of Gronings on the basis of *Klunderloa*, a website with texts for primary school children,[3] as well as the Reker dictionary of Gronings (Reker, 1998). The initial lexicon contained 35,012 unique words. This increased the chance of the program correctly recognising a word it was not certain about and made the task of the volunteers easier. After the initial OCR step, the program presented the volunteers with all words of which it was not certain whether they were recognised correctly. The volunteers then had to compare the text as recognised by the program to the scan, and correct the text if necessary. If a word had not been encountered by the program before, this was also indicated and volunteers were presented with the opportunity to add this word to the lexicon in order to facilitate recognition in the future. As the goal of the corpus was to serve as an accurate representation of all forms of written Gronings, no alterations to the original texts were made. As the spelling of Gronings shows substantial variation diachronically, between variants, and also between authors, it is impossible to make an objective distinction between typing and spelling errors on the one hand, and intentional ‘non-standard’ forms meant to reflect differences in pronunciation on the other hand. Therefore, volunteers were explicitly instructed to only perform corrections on the texts if the OCR output did not match the text in the scan that they were presented with, and to leave in all other ‘errors’ they might perceive. Some of the texts were not suitable for OCR, as they used non-standard fonts (for example to resemble cursive handwriting), because the text was overlaid on a background image where parts of the image could be confused for text (such as drawings) or (especially for the older texts) because the quality of the paper and/or printing was poor. These texts (<5% of the total) were taken

[3] www.klunderloa.nl

out of the pipeline and stored in a separate folder for potential later correction, as it would take the volunteers too much time to transcribe these texts manually.

3.3. Adding Metadata

After the OCR results were checked, the files were transported to a website that allowed volunteers to do both a final check of the text and to add metadata. Some volunteers preferred to conduct this step themselves for each text they did the OCR check for, and some only did one of two steps. Both of these options worked well. For this step, we designed a custom application that allowed volunteers to view (1) the scan, (2) the (editable) text as produced in the OCR step, and (3) forms through which they could add the metadata. The metadata that volunteers were asked to add consisted of two parts: metadata about the whole text, and metadata about different parts of the text. The metadata about the whole text consisted of editor, title, source type (book, journal, newspaper, website), series, year, number, place of publication, publisher, edition or printing, website, date of consulting website, and comments. The metadata about different parts of the text consisted of author, title, genre (prose, poetry), first language variant (normally Gronings), second language variant (if another language variant was is used as well), and comments. The metadata was partly found in the sources themselves, and partly needed to be looked up online or in reference works. If the data were available through the University Library, the form fields were filled in automatically with those data.

4. Adding Lemmas and Part-of-Speech Tags

4.1. Lemmatisation

We developed a lemmatiser which lemmatises tokens in Gronings to lemmas in Dutch. Assigning Dutch lemmas to tokens in texts that are written in Gronings is important for two reasons. It (1) allows the user to search the corpus in both Gronings (via the tokens) and Dutch (via the lemmas), and (2) regional, morphological and spelling variants of the same word are 'linked' in this way. For example, if a user searches by using the Dutch word *huis* 'house', sentences with all occurring Gronings variants are found: *hoes*, *huus*, *hoeske*, *huusie*, etc, representing respectively two different regional forms of the base word and two different regional forms of the diminutive. If the user searches for the Groningen word *hoes*, it is also possible to not only find sentences that include the exact word *hoes*, but also sentences that include *huus*, *hoeske* and *huusie*. In this way, forms of regional, diachronic and idiosyncratic spelling variation are preserved and made accessible in the corpus.

To be able to lemmatise automatically, a lemmatiser had to be trained on the basis of a training corpus. Our training corpus consisted of six texts in Gronings, containing 109,765 tokens, 93,739 words and 6,513 sentences in total. When assigning the lemmas, a Dutch cognate was chosen whenever possible. If there was no cognate in Dutch for the Gronings word, a non-cognate was chosen. This training corpus was manually created as a part of our project. We estimate that the creation of this corpus, including the training of a student assistant, took 150 hours.

For lemmatisation, we trained the PIE (Manjavacas et al., 2019) lemmatiser. We chose this lemmatiser as it is robust in the presence of much language variation, as is the case for our corpus. On the one hand there is regional and diachronic variation, and on the other hand authors use different spellings. The accuracy of our model was determined to be 89% through 10-fold cross validation. A visual inspection suggests that a substantial portion of the errors are cases where the model generates a Dutch-sounding cognate that is not commonly used, while the word was previously annotated in the training corpus with a non-cognate. When no cognate in Dutch is present at all and the word was not included in the training corpus, the lemma is derived from or identical to the token. We do not consider this a problem since different variants of Gronings still normalise to the same (pseudo-)Dutch lemma, and this is the primary goal of the lemmatisation process (although in cases where no cognate is present, this can mean that the word is not findable through the Dutch lemma).

4.2. Part-of-Speech Tags

Assigning part-of-speech (POS) tags to the words is important because some words in Gronings – just like some Dutch words – belong to a different part of speech depending on the context in which they appear. For example, there are three POS-tags for the word *aal* (an adverb when the meaning is 'constantly', a pronoun when the meaning is 'everyone' and a noun when the meaning is 'the universe'). Consequently, in order to search the corpus for appropriate sentences containing the word *aal*, one needs to specify the part of speech.

We automatically added POS-tags to our corpus with a BERTje-based language model. BERTje is a general language model for Dutch (de Vries et al., 2019). This model was trained for Dutch POS tagging, based on training data from the Universal Dependencies project (de Marneffe et al., 2021). Additionally, the model was adapted to work with words in Gronings through a multi-step adaption process. In this process, the model was fine-tuned for POS tagging in Dutch, and adapted to Gronings using unlabeled data (de Vries et al., 2021) and reached an accuracy of 92% on the unseen Gronings test set. Since the POS tagging model is trained cross-lingually using Dutch training data, there should not be a bias towards a specific Gronings variant, but the model might perform better for variants that are more similar to Dutch.

POS tags are useful discriminators for semantic disambiguation (Wilks and Stevenson, 1996). However, they are not enough to fully disambiguate a text. For example, *bank* can be a financial institute or the edge of a river. In both cases *bank* is tagged as a noun. Therefore, a useful refinement would be to assign the appropriate sense to each occurrence of the word in a given context, a process known as sense tagging (Wilks and Stevenson, 1997). In order to train a sense tagger, you need to annotate a training corpus with word senses, a task that may be time-consuming. Due to the limitations of our project, this has not been done yet, but will be useful future work.

4.3. XML

The final result consists of texts in XML format that contain the metadata and in which the words are annotated with their lemmas and POS tags. These texts are suitable to be searched by the BlackLab corpus search engine (de Does et al., 2017). BlackLab is a corpus retrieval engine built on top of Apache Lucene and used by the newly developed corpus search interface in WoordWaark.

The interface offers four search options allowing for varying search query complexity: simple, extended, advanced, and expert. The basic search option enables the user to search for specific words, while the advanced options allow for more complicated search queries involving partial words, lemmas, and POS tags. The input provided by the user is converted into CQL (Corpus Query Language), a query language used by BlackLab to allow users to retrieve information from the available corpora. The server's response is presented in the form of a table, with the matching word(s) displayed together with its surrounding context. Those words are clickable and take the user to the corresponding lemma in the dictionary. Additionally, details concerning each text in which the search term appears, such as the title and author, can be easily viewed.

5. Other Considerations & Lessons Learned

One of the main difficulties we expected in building the corpus was having to account for the substantial variation that would be present in the data. However, by using PIE and a manually annotated dataset for lemmatisation together with an adapted version of BERTje, we still achieved results that are sufficiently accurate for a general audience and that would greatly aid researchers in providing a first crude annotation of the data. As manual tagging and lemmatisation would not be feasible for corpora of this size, we think this method is suitable for other languages as well. It is important to note, however, that the effectiveness of this approach is dependent on the presence of linguistic resources from a closely related (standardised) higher-resource language (de Vries et al., 2021).

Another recommendation for similar projects in the future concerns the use of volunteers. Although our volunteers were highly intrinsically motivated to partake in this project, they indicated that it was sometimes demotivating that the work they did was very individual. Because of the COVID-19 pandemic, we were unfortunately not able to organise many activities or (informative) gatherings for the volunteers, but would recommend this for similar projects in the future. It was evident, once this was again possible, that the volunteers enjoyed seeing the results of their work illustrated through presentations about WoordWaark and research conducted on the corpus at the university.

6. Conclusion

Both the infrastructure designed for this project and the lessons learned from it may be useful for other under-resourced languages with internal variation for which the construction of a written corpus would be desirable. The current paper has demonstrated a method in which a combination of volunteer work and automation creates an efficient pipeline for converting printed texts to annotated sentences which are potentially useful for a general audience and researchers. Furthermore, we have demonstrated how resources from a larger related language (Dutch) can be usefully employed for a (related) low-resource variety and how challenges concerning spelling variation can be circumvented while preserving the variation in the corpus. As the infrastructure of the corpus was designed to be used by other languages as well, a pilot is currently underway in which the infrastructure will be used for Bildts, another minority language variety that is spoken in the Netherlands. Furthermore, the complete pipeline, manuals for software and coaching volunteers as well as the software designed for the project are available in the project's GitHub repository.[4]

[4] github.com/woordwaark/Spotlight-pipeline

7. Acknowledgements

We would like to thank the volunteers who helped with the creation of the corpus. In addition, we are grateful for the financial contribution supporting the improvement of WoordWaark by a foundation which wishes to remain anonymous.

8. Ethical Considerations

One of the main ethical considerations we encountered during the construction of our corpus is that it can be difficult to adequately take into account the interests of the two target audiences that might be using the corpus. As the corpus should both be usable for academic research and for a general audience trying to gain insight in the usage of specific words, some conflicts arose in which sentences were appropriate to include. All material from the texts that was in principle usable was included in the corpus, which meant that there were also sentences containing racist, sexist, homophobic and other offensive language. Although it is necessary to include these sentences for linguistic research, they are not appropriate to present to a general audience as examples of how other (inoffensive) words are used in the language. Therefore, we constructed a list of words that caused sentences containing one or more of these words to not be shown as illustrations of the use of a different (inoffensive) word in that sentence when using the basic search functionality. In case someone would deliberately search for an offensive term, the sentences containing these terms are shown, however. We feel that this approach best combines the interests of both researchers and a general audience, as the sentences containing offensive terms are still accessible using the more complex searching functionality used by researchers, but would not be presented as examples that could be seen as normative to a general audience.

9. Bibliographical References

Annelie Ädel. 2020. Corpus compilation. In Magali Paquot and Stefan Th. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pages 3–24. Springer International Publishing, Cham.

Henk Bloemhoff. 2005. *Taaltelling Nedersaksisch. Een enquête naar het gebruik en de beheersing van het Nedersaksisch in Nederland*. Stichting Sasland, Groningen.

Jess de Does, Jan Niestadt, and Katrien Depuydt. 2017. Creating research environments with BlackLab. In Jan van Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, pages 245–257. Ubiquity Press, London.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model.

ECRML. 1998. Europees Handvest voor Regionale Talen of Talen van Minderheden, Straatsburg, 05-11-1992.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Musat, and Andreas Fischer. 2019. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Alice Millour and Karën Fort. 2020. Text corpora and the challenge of newly written languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 111–120, Marseille, France. European Language Resources association.

Siemon Reker. 1984. *Groninger spelling. Handleiding voor het lezen en schrijven van Groninger teksten*. Stichting ’t Grunneger bouk, Haren.

Siemon Reker. 1998. *Zakwoordenboek Gronings-Nederlands, Nederlands-Gronings*. Staalboek, Veendam.

Bilal Tahir and Muhammad Amir Mehmood. 2021. Corpulyzer: A novel framework for building low resource language corpora. *IEEE Access*, 9:8546–8563.

Kornelis Ter Laan. 1947. *Humor in Grunnegerlaand*. Strengholt, Amsterdam.

Jennifer Tracey, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, and Neil Kuster. 2019. Corpus building for low resource languages in the DARPA LORELEI program. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55, Dublin, Ireland. European Association for Machine Translation.

Arjen Versloot. 2020. Streektaaldood in de Lage Landen. *Taal en Tongval*, 72(1):7–16.

Yorick Wilks and Mark Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield, Sheffield.

Yorick Wilks and Mark Stevenson. 1997. Sense tagging: Semantic tagging with a lexicon. In *Tagging Text with Lexical Semantics: Why, What, and How?*