Constructing SABeD: A Spoken Academic Belgian Dutch Corpus

Jolien Mathysen
KU Leuven, Belgium
jolien.mathysen@kuleuven.be

Vincent Vandeghinste
Instituut voor de Nederlandse Taal, The Netherlands
vincent.vandeghinste@ivdnt.org
KU Leuven, Belgium

Elke Peters
KU Leuven, Belgium
elke.peters@kuleuven.be

Patrick Wambacq
KU Leuven, Belgium
patrick.wambacq@kuleuven.be

Abstract

We present the Spoken Academic Belgian Dutch (SABeD) corpus and a description of its construction. It was compiled from selected first bachelor academic lectures in higher education institutions in Flanders, as students indicate that the language used in such lectures is one of the hurdles for comprehension and academic success. We first applied speech recognition on these lectures, and then applied manual utterance segmentation and manual correction of the automated transcription. A filtered version of the resulting transcriptions was automatically punctuated and linguistically annotated with CLARIN tools and is currently available for search in the Autosearch online corpus query environment. The manual transcriptions and the ELAN files with the final annotation will soon be made available to the research community for download in the CLARIN infrastructure at http://hdl.handle.net/10032/tm-a2-w4.

1 Introduction

In higher education, students are confronted with academic language use, with which they are often not familiar. Since academic language skills are a necessary condition for study success, higher education institutions in Flanders and the Netherlands focus on language support for students. In many institutions, these efforts evolved into formal, embedded language policies, but research into their implementation is limited (Bonne & Casteleyn, 2022). The number of international students pursuing higher education in Flanders is estimated around 2500 per year (Deygers & Malone, 2019). Research (Deygers, 2017; Deygers et al., 2017) shows that Dutch language learners struggle with academic spoken Dutch, even when they passed the university entrance language tests, ITNA (*Interuniversitaire Taaltest Nederlands voor Anderstaligen*; Interuniversity Test of Dutch for Speakers of Other Languages)[1] or CNaVT (*Certificaat Nederlands als Vreemde Taal*; Certificate of Dutch as a Foreign Language)[2]. These tests are required for international students wishing to follow programmes taught in Dutch in higher education in Flanders. Although academic listening is part of the test, learners have indicated that the listening tasks used in the tests are easier than the actual lectures (Deygers et al., 2018). One reason for the discrepancy is that the linguistic features of the listening task in the test have not been empirically validated because of the lack of a corpus of spoken academic Dutch.

One of the key predictors of success in university entrance language tests (Heeren et al., 2020; Trenkic & Warmington, 2019) is learners' vocabulary knowledge. At the same time, academic vocabulary

[1] https://www.itna.be/
[2] https://cnavt.org/

has also been put forward as one of the main challenges that university students face when listening to academic lectures. Research into English has shown that listening comprehension increases with lexical coverage, or the percentage of known words in a text (Durbahn et al., 2020; Van Zeeland & Schmitt, 2013), and that vocabulary knowledge is a strong predictor of academic listening comprehension (Wallace, 2020). More concretely, learners should be familiar with 95% to 98% of the running words in spoken academic English to reach reasonable to very high comprehension. This means that, to reach detailed comprehension of lectures, learners need to know high frequency vocabulary, as well as the 3,000-9,000 most frequent word-families (plus proper nouns and marginal words) in English respectively (Dang, 2022; Dang & Webb, 2014; Stæhr, 2009).

This was one of the main reasons for starting the SABeD (Spoken Academic Belgian Dutch) project. The principal aim of this project was building a multimodal corpus of spoken academic Belgian Dutch, which consists of (transcripts and audio/video recordings of) academic lectures[3]. Besides, lectures can be said to constitute the dominant form of instruction for students entering higher education in Flanders. This means that they must process new subject matter in a register they may not yet fully master. This is particularly problematic, considering that research has shown a link between academic language skills and success in higher education (Heeren et al., 2020, 2021; Milton & Treffers-Daller, 2013; Trenkic & Warmington, 2019). The compilation of a Spoken Academic Belgian Dutch corpus is, thus, warranted. Especially since earlier research (Dang et al., 2017) has already demonstrated the usefulness of such corpora in analysing the lexical demands of academic lectures and in developing/validating vocabulary learning materials that can help students reach the required lexical coverage levels for detailed comprehension of those lectures. Moreover, such corpus-based learning materials may also raise awareness among both teachers and students about their academic language use and struggles (Dang, 2022; Dang et al., 2021; Nation, 2016; Uchihara & Webb, 2022). Due to the covid pandemic, recorded video lectures are available in abundance. The SABeD corpus contains a mix of written-to-be-read and spontaneous spoken language.[4] This paper presents how the corpus was created.

Section 2 describes related work, Section 3 describes how the corpus was compiled, and Section 4 describes the semi-automatic transcription process, consisting of a phase of automated speech recognition (Section 4.1) and a phase of manual post-editing (Section 4.2). Section 5 describes how the manual transcriptions were filtered and postprocessed using CLARIN tools and how the corpus is made available to specific researchers, again using CLARIN tools. Section 6 draws conclusions and sketches the plans of using the corpus for the creation of an academic vocabulary list and empirical validation of a new university entrance vocabulary test.

2 Related Work

Some of the most notable corpora featuring spoken academic language currently in existence include the MICASE (*Michigan Corpus of Academic Spoken English*) corpus (Simpson et al., 2002), T2K-SWAL (*TOEFL 2000 Spoken and Written Academic Language*) corpus (Biber et al., 2002), BASE (*British Academic Spoken English*) corpus (Thompson & Nesi, 2001) and the CGN (*Spoken Dutch Corpus*) (Oostdijk et al., 2002). We will briefly discuss these four corpora because of their prominence within the field, and the standards they provided for designing the SABeD corpus in terms of balance and representativeness decisions.

[3] The other aims of the project are investigating the effectiveness of speech technology for automatic transcription of spoken texts, developing a word frequency list of spoken academic Belgian Dutch and creating a vocabulary test of spoken academic Belgian Dutch, but these fall outside of the main focus of this paper.

[4] In analogy with CGN (Oostdijk et al., 2002), we consider our corpus a *spoken corpus*. The initial collection of 972 recordings did include a number of pre-recorded lectures and knowledge clips, in which the lecturers had prepared and read out their text. However, in the selection of the 200 recordings for the final corpus, live lectures taught on campus were given priority. Additionally, other spoken corpora, such as the CGN also feature a variety of (semi-)structured instances of speech (e.g. interview, news bulletins, masses, formal speeches and even recited texts) (cf. https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_Dutch/topics/overview.htm#inleiding). Apart from that, other spoken academic corpora, among other the MICASE corpus (Simpson et al., 2002), also conceptualise academic speech in a broad sense as any speech occurring in an academic setting. Consequently, they also tend to contain (semi-)structured instances of speech, such as presentations or dissertation defenses.

The MICASE corpus comprises 152 different academic speech events (i.e. small and large lectures, discussion and lab sections, seminars, student presentations, advising sessions, colloquia, dissertation defenses, interviews, meetings, office hours, service encounters, study groups, tours and tutorials) from the University of Michigan, amounting to nearly 200 hours or approximately 1.7 million words worth of transcriptions. It spans four major academic divisions (i.e. Humanities and Arts, Social Sciences and Education, Physical Sciences and Engineering, and Biological and Health Sciences) and also features discourse and text linguistic annotation (Simpson et al., 2002). The spoken component of the T2K-SWAL corpus includes 1.7 million words recorded at four different American universities. The largest part of this (1.2 million words) was taken from 176 class sessions, while the remaining 50,000 came from office hours (Biber et al., 2002). BASE consists of 160 lectures and 40 seminars recorded at the University of Warwick and the University of Reading between 2000 and 2005. This 1,186,290 token corpus was compiled from four disciplinary sub-corpora: Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Sciences. Except for Physical Sciences, each sub-corpus contains 40 lectures and 10 seminars (Thompson & Nesi, 2001). CGN is a balanced corpus with several variants of spoken Dutch (from read-aloud text to spontaneous conversations, from Belgium and the Netherlands), and which contains 30,917 words from university lectures (Oostdijk et al., 2002). Apart from these four corpora, clearly, other less-known and/or smaller spoken academic corpora have been created for English, as well as for other (non-Indo-European) languages. However, a domain-specific spoken corpus like the one we introduce in this paper was until recently not available for (Belgian) Dutch.

3 Corpus Compilation

In the corpus compilation stage, we selected academic lectures from both KU Leuven and Ghent University, because these constitute the predominant form of instruction in higher education institutions in Flanders, especially in the first bachelor year. Our corpus does not include university college lectures, because it was not possible to collect as many of them as of the university lectures. In addition, the university college lectures that we did manage to collect were predominantly associated with the social sciences. As such, including these recordings in the corpus without obtaining any more university college lectures first would have resulted in an unbalanced corpus.

For the purpose of our corpus, lectures are defined as instructional discourse presented before an audience of at least 40 students, in which the lecturer is the dominant speaker and the level of interactivity is modest to low. This is in line with the distinction made by the MICASE (Simpson et al., 2002) corpus between small and large lectures, for which the cut-off point was also 40 students. Furthermore, there are several reasons validating a corpus specifically oriented toward this kind of prototypical lecture. First, one should consider if and to what degree the language of lectures is influenced by the size of their audiences (e.g. frequency and nature of the interactions between lecturer and students). In this regard, it should also be taken into account that both native speakers and foreign learners of Dutch specifically indicated the language used in Flemish lectures as one of the hurdles for comprehension and academic success (Deygers, 2017; Deygers et al., 2017). In addition, research for English has also found that lectures are more (lexically) demanding as opposed to e.g. labs and tutorials (Dang et al., 2021).

Academic Division	**No. of Videos**	**No. of Speakers**	**Full Video length**	**Transcribed length**	**Tokens (raw)**	**Tokens (clean)**
Biological and Health Sciences	50	5	56:20:04	25 hrs	201,122	190,943
Physical Sciences and Engineering	50	11	75:16:55	25 hrs	207,398	200,883
Humanities and Arts	50	9	61:54:18	25 hrs	227,360	215,374
Social Sciences and Education	50	6	99:55:52	25 hrs	267,279	254,487
Total	200	31	293:24:09	100 hrs	903,159	861,787

Table 1: Corpus size over the different academic divisions. Video length is shown in `HH:MM:SS` (Hours:Minutes:Seconds format) and represents the lengths of entire lectures.

We specifically chose first year bachelor lectures, because they also constitute the first encounter of the target group (i.e. Flemish first year bachelor students and international students commencing university education in Belgian-Dutch) of our corpus with spoken academic Dutch. In this regard, it is particularly important to take into account the primary pedagogical goal of the SABeD project and corpus, i.e. developing learning materials for students entering Flemish higher education. As such, these lectures make up a solid base for our corpus compilation, especially considering that we cannot be certain if and to what degree the language of lectures in later bachelor and master years differs from that in the first bachelor year.

To ensure that the corpus is both representative and has sufficient power to make statistical inferences, lectures from a considerable number of lecturers need to be included (Biber, 1993). However, the selection of lectures for the transcription stage was impeded by the fact that, due to technical reasons, it was not possible to automatically download lectures from the video platforms used by Flemish universities. All lecturers/professors had to be contacted individually and all lectures had to be downloaded and processed one-by-one before they could be added to the corpus. Informed consent was obtained and metadata was collected (e.g. speaker data such as age, gender, teaching experience, place of birth). Eventually, we obtained 972 recordings, out of which we selected 200 lectures to undergo manual post-editing. This is still more lectures than in other existing corpora of spoken academic English such as the MICASE corpus (62 lectures; (Simpson et al., 2002)), T2K-SWAL corpus (176 lectures; (Biber et al., 2002)) or BASE (160 lectures; Thompson and Nesi, 2001). In terms of tokens (cf. Table 1), our corpus is smaller than the corpora mentioned in Section 2, but this is mainly due to the fact that we chose to only manually transcribe 30 minutes per lecture instead of transcribing a considerably smaller number of entire lectures. Only the 30-minute parts of lectures that had undergone manual post-editing were included in our base corpus to ensure a well-balanced corpus. However, the parts of the lectures that were transcribed strictly automatically will also be included at a later stage. Additionally, the remaining 772 videos that were not selected to be part of the initial base corpus will also be processed and added later, thus ensuring that the corpus will still grow significantly in size.

	Biological and Health Sciences	**Physical Sciences and Engineering**	**Humanities and Arts**	**Social Sciences and Education**
Anatomy	23	0	0	0
Archaeology, Art, History, Philosophy	0	0	10	0
Biochemistry	14	0	0	0
Chemistry	0	17	0	0
Dutch linguistics	0	0	18	0
Economics and Law	0	0	0	16
Electronics and Programming	0	5	0	0
General Linguistics	0	0	22	0
Genetics	13	0	0	0
Maths	0	13	0	0
Physics	0	15	0	0
Psychology	0	0	0	18
Research in Social Sciences	0	0	0	16
Total	50	50	50	50

Table 2: No. of lectures per discipline and academic division

Due to us only manually transcribing 30 minutes per lecture, the 200 selected lectures had to have a length of at least 30 minutes. However, this criterium was also to a lesser extent dictated by the existing data and circumstances. First, the prevalence and use of shorter lecture recordings (between 30 minutes and 1 hour) seems to have increased since the covid-19 pandemic and thus in the data that we collected as well. Secondly, including these shorter lectures also led to greater variation in terms of speakers in the corpus. Following the example of the MICASE (Simpson et al., 2002) and BASE (Thompson & Nesi, 2001) corpora, the selection of lectures was further informed by academic division (Biological and Health Sciences, Humanities and Arts, Physical Sciences and Engineering, Social Sciences and Education). Table 1 presents the size of each of the academic divisions in the corpus. Generally, the academic division was attributed to a lecture by checking the study programme in which the lecture had been taught. In ambiguous cases, for instance a science course being taught as part of the archaeology programme, the research unit and department of the lecturer determined the allocation of the academic division. Within each academic division, a broad range of disciplines is covered. We aimed to compose a corpus that

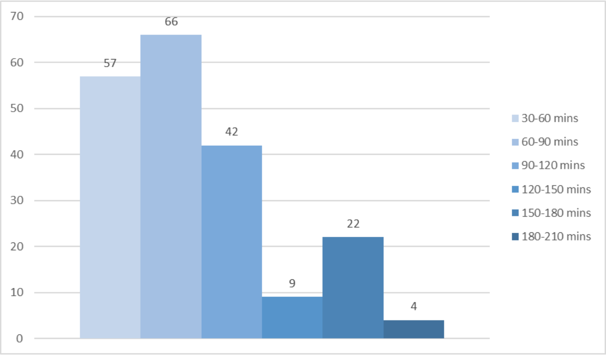


Figure 1: No. of lectures in the corpus according to their length in minutes

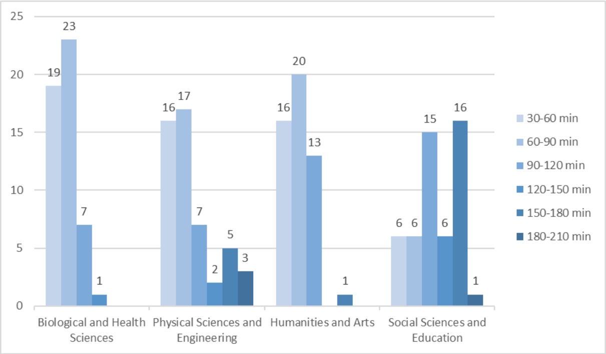


Figure 2: No. of lectures in the corpus according to their length in minutes per academic division

is sufficiently representative for the purpose of composing a word list of spoken academic Dutch. At the same time, the corpus should contain sufficient data to obtain standards for more specific academic divisions. Consequently, each of the aforementioned academic divisions was equally represented, thus amounting to 50 lectures per division. A distribution of the lengths of the videos is shown in Figure 1. A more detailed overview of the range of disciplines represented by the videos in the corpus per scientific division is given in Table 2, and a distribution of the length of the videos per academic division is shown in Figure 2. Table 2 was obtained by grouping together the videos of courses in the corpus that covered identical, similar or related subjects within each scientific division. This was done to be able to provide additional insight into the disciplines that are incorporated while also considering the privacy of lecturers and GDPR (General Data Protection Regulation) by not sharing the actual course names. Finally, we also collected about 800,000 tokens of written course materials, to increase the accuracy and performance of the speech recognition system (cf. section 4.1).

4 Transcription

As mentioned before, we only manually transcribed 30 minutes per lecture, instead of transcribing a considerably smaller number of entire lectures. More specifically, we transcribed only the first 25 and last 5 minutes of the lectures. The main benefit of this practice is that any differences between institutes and disciplines in length of lectures are eliminated, ensuring a well-balanced corpus. Moreover, transcribing the introductory parts and conclusions of the lectures ensured that the corpus is sufficiently representative of the more general spoken academic language spanning across different divisions. In these parts one could, in fact, expect a higher occurrence of general academic language and vocabulary than in the middle parts which could contain more subject-specific technical language and vocabulary. However, by also transcribing portions of the lectures that stretch beyond their introductions, the corpus also still captures data on more specific academic divisions and disciplines. This distinction between general academic and technical language/vocabulary is based on that of Nation, 2016.

4.1 Automated Speech Recognition

Speech recognition was performed with an ASR system tuned for Belgian Dutch (Van Dyck et al., 2021), which is Kaldi-based (Povey et al., 2011). The Kaldi toolkit makes use of state-of-the-art deep neural networks. The new acoustic model was trained on data from the Spoken Dutch Corpus (Oostdijk et al., 2002) and tested using the N-best benchmark (Kessens & van Leeuwen, 2007). The output of the ASR system consists of `ctm` files,[5] which contain time stamps for each recognized word and a confidence level for each word. An alternative for this system would be to use the recently emerging end-to-end systems, but these do not allow independent training of language models, which is one of the additional goals of the SABeD project. The additional text material from textbooks and course materials will be used to improve the lexicon and language model independently from the acoustic model. This latter effort is still ongoing. The *raw ASR* output was inserted into an ELAN file as a separate tier. ELAN (Wittenburg et al., 2006) is well known software for linguistic annotation of audio-visual material.

4.2 Manual Post-Editing

Manual post-editing was performed in two stages. Section 4.2.1 describes utterance segmentation and Section 4.2.2 describes manual correction of the speech recognition results.

4.2.1 Utterance Segmentation

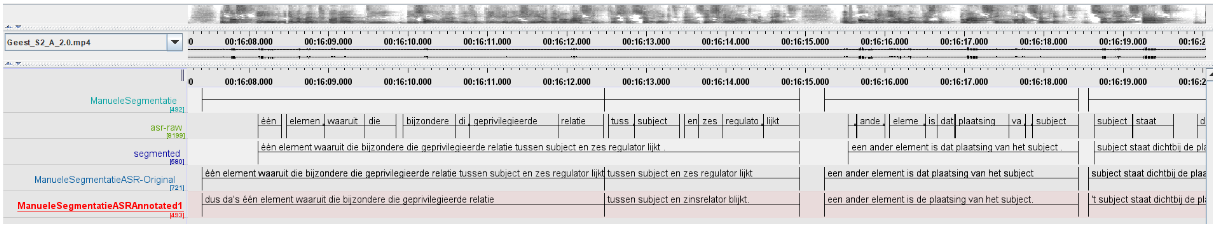
A first step in manual post-editing consisted of applying utterance segmentation. This was done to make post-editing easier and faster. If we would not have applied segmentation, ASR correction would necessarily have needed to be performed at the word level, which would have been cumbersome as ASR errors can span over word boundaries, requiring annotators to not only correct the transcript, but also to manually manipulate the time stamps. This segmentation annotation was registered in a separate ELAN tier. It entailed the placement of boundaries which indicate the start and ending of a piece of transcription in the audio signal. The unit which demarcates these boundaries is the chunk, which is defined as a speech fragment which lasts about 2-3 seconds and which is delimited on both sides by a (short) audible and visible pause. Chunks can, but do not need to, correspond to sentences or phrases.

Shorter pauses (nearly always shorter than 1 second) that occurred within a chunk were ignored. As such, chunks which lasted less than 2-3 seconds were only possible if we came across a speech fragment of less than 2-3 seconds that was, on both sides, clearly separated from any other fragments by a long pause (i.e. of more than 3 seconds). Chunks longer than 3 seconds were allowed (e.g. multiple co- or subordinated clauses, long enumerations). However, if chunks lasted longer than 6 seconds, they were split up in front of a conjunction or where a comma would appear in written language. All in all, the length of chunks was highly dependent on the pace of the speaker under scrutiny, with slower speakers generally yielding less and longer chunks and faster speakers usually requiring more and shorter ones. An absolute criterion for the segmentation consisted in keeping compounds within the same chunk.

Any student interactions or substantial background noises (e.g. an opening door) were also isolated in segments and tagged in a separate ELAN tier. This way students' voices could be easily removed later to respect their privacy and GDPR, as we could not identify them nor obtain their informed consent. Noises could be taken into consideration when influencing the flow of speech. If a noise persisted in the background throughout all or most part of the recording, no segment was created and this was signalled by adding a comment to the ELAN file. The *raw ASR* tier is then combined with the *manual segmentation* tier into a *segmented ASR* tier[6].

[5] `ctm` stands for time-marked conversation file.

[6] Privacy and GDPR inhibit us from sharing the contents of both the *raw ASR* tier and the *segmented ASR* tier, since the automatic transcriptions may contain personal data (e.g. names of lecturers, courses, student names) or speech from students of whom we did not get informed consent. All this information was removed in a later stage of the transcription process.



ManueleSegmentatie [492]				
asr-raw [8199]	één elemen waaruit die bijzondere di geprivilegieerde relatie	tuss subject en zes regulato lijkt	ande eleme is dat plaatsing va subject	subject staat d
segmented [580]	één element waaruit die bijzondere die geprivilegieerde relatie tussen subject en zes regulator lijkt .		een ander element is dat plaatsing van het subject .	subject staat dichtbij de pla
ManueleSegmentatieASR-Original [721]	één element waaruit die bijzondere die geprivilegieerde relatie tussen subject en zes regulator lijkt	tussen subject en zes regulator lijkt	een ander element is dat plaatsing van het subject	subject staat dichtbij de plaa
ManueleSegmentatieASRAnnotated1 [493]	dus da's één element waaruit die bijzondere die geprivilegieerde relatie	tussen subject en zinsrelator blijkt.	een ander element is de plaatsing van het subject.	't subject staat dichtbij de pla

Figure 3: ELAN screenshot showing the different tiers for manual post-editing

4.2.2 Correcting the Automated Transcriptions

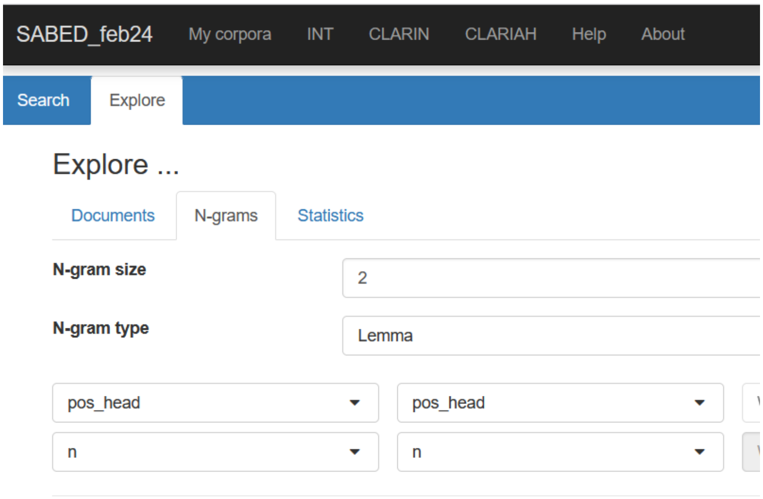
The second step consisted of correcting the automated transcription at the segment level. Annotators manually corrected the *segmented ASR* tier using a transcription protocol that was based on that of the CGN (Oostdijk et al., 2002) and put the correction into the *manual transcription* tier. More specifically, there were two phases involved in the manual correction of the transcriptions. During the first phase, transcriptions were corrected at the orthographic level. This entailed that the spelling was standardised and punctuation was added. Interjections that did not constitute words themselves (e.g. *ah, hè, uhm*) and words from languages other than Dutch (e.g. Latin medical terms or English book titles, but not loanwords that have been accepted into Dutch) were annotated using designated codes. Students' speech was cut out because it was too difficult to track them down and get their consent. Personal data (e.g. names of lecturers, students, courses) were also anonymised to uphold privacy and GDPR. During the second phase, transcriptions were checked at the acoustic level. This included the annotation of reductions (e.g. *'k moet hier weg*; I got to get out of here or *da's niet waar*; that's not true), dialect words (e.g. *talloor*; dialect term for plate), slips of the tongue, aborted words and sentences, unintelligible pieces of speech, and noises made by the speakers (e.g. coughing or sneezing). The number of files processed and the token count (using the linux `wc` command) on the extracted *manual transcription* tier are presented in Table 1.

5 Processing with CLARIN Tools

The ELAN files containing the manual transcriptions and other annotation layers are available for download at https://hdl.handle.net/10032/tm-a2-w4. When extracting the transcription for further processing and extraction of frequency lists, we removed the special codes that were annotated in the manual transcription, such as tokens indicating hesitation, or non-understandable audio pieces, and special indicators marking dialect and foreign words. Resulting token counts after such cleaning are presented in Table 1. Note that these codes were included in the manual corpus annotation as such annotations may be relevant for other types of research related to spoken language or as training data for speech recognition models.

The corpus was then processed with the Full Stop Punctuation Inserter for Dutch (Vandeghinste & Guhr, 2023), which was built specifically for projects like SABeD. The resulting corpus was then processed using the CLARIN tool Frog (van den Bosch et al., 2007),[7], an NLP processing toolkit for Dutch. This resulted in the following analyses: tokenization, part-of-speech tagging, lemmatization, morphological segmentation, dependency parsing and named entity labeling. Frog outputs FoLiA format (van Gompel & Reynaert, 2013), an XML format made for linguistic annotations and CoNLL type tab separated files, containing the same information.

[7] https://webservices.cls.ru.nl/frog



Explore ...

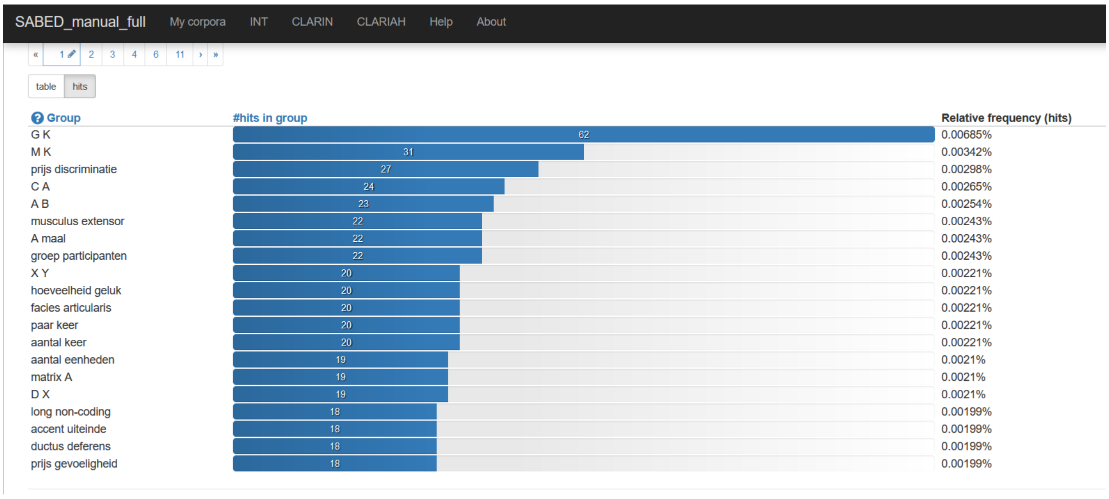
N-gram size: 2

N-gram type: Lemma

pos_head: n

pos_head: n

Figure 4: Query for generating bigrams of lemmas of nouns in Autosearch



Group	#hits in group	Relative frequency (hits)
G K	62	0.00685%
M K	31	0.00342%
prijs discriminatie	27	0.00298%
C A	24	0.00265%
A B	23	0.00254%
musculus extensor	22	0.00243%
A maal	22	0.00243%
groep participanten	22	0.00243%
X Y	20	0.00221%
hoeveelheid geluk	20	0.00221%
facies articularis	20	0.00221%
paar keer	20	0.00221%
aantal keer	20	0.00221%
aantal eenheden	19	0.0021%
matrix A	19	0.0021%
D X	19	0.0021%
long non-coding	18	0.00199%
accent uiteinde	18	0.00199%
ductus deferens	18	0.00199%
prijs gevoeligheid	18	0.00199%

Figure 5: First page of results from querying lemma bigrams for nouns.

These FoLiA data have been uploaded to the CLARIN Autosearch engine[8] for annotated corpora, making the corpus now searchable with Corpus Query Language (CQL) and sharable with other CLARIN users. Figure 4 shows the query to generate lemma bigrams of nouns in the corpus, and Figure 5 shows the results page. Results are also downloadable as csv files. While it is clear that the results are not always informative, the Autosearch interface to the corpus provides easy ways of exploring the corpus, and the queries can easily be made more refined. Autosearch can also be used as a concordancer.

6 Conclusions and Future Work

We have presented the corpus compilation efforts for a corpus of spoken academic Belgian Dutch. Because we prioritised manual post-editing of parts of a limited selection of lectures and balancing, our corpus is currently limited in token size (cf. Table 1) compared to other similar corpora. However, the strictly automatically transcribed parts of this selection of lectures, as well as 772 videos that were not part of the base corpus, will still be added at a later stage. This will significantly expand the corpus' size

[8] Autosearch is a version of Blacklab (de Does et al., 2017), which works with customizable corpora. http://hdl.handle.net/10032/tm-a2-c5

in the future. Once all data is processed, combining the metadata with the linguistic annotations in TEI format will allow even more fine-grained querying of the corpus, not only on linguistic criteria but also on metadata criteria, and the corpus will be made available in a Blacklab (de Does et al., 2017) corpus query engine for all CLARIN users. Even so, we still acknowledge that variation in terms of speakers and disciplines was one of the main challenges and concerns in the construction of this corpus. Additionally, we would like to point out that the matter of anonymising multimedia materials (with regard to personal data such as names of lectures, course names and students' speech), specifically audio and video, should not be taken lightly. Nevertheless, we aim to make the video and audio recordings of the lectures available to the research community as well once this process has been completed.

Concerning automatic speech recognition we can assert that while ASR speeds up the manual transcription, it is clear that a general domain ASR system does not contain a specialized vocabulary like the one that is being used in academic lectures, and therefore tuning the vocabulary and language model of the ASR system towards the specific domains is expected to greatly improve ASR accuracy and reduce post-editing effort, which should result in a speedier post-editing process.

Even though validation of a new word list and vocabulary test for academic Belgian Dutch was one of the main reasons for collecting the corpus, this is still future work. The development of an academic spoken word list will be based on the frequency and range of the words in the corpus (Dang et al., 2017; Szudarski, 2017) with the lemma as counting unit. This functionality is included in the Blacklab environment. To determine which words can be considered academic words, the frequency list will be compared to the word list of Tiberius and Schoonheim, 2013. Words not occurring in that list are potential candidates for the spoken academic vocabulary list, depending on their frequency and distribution in the corpus. We will distinguish proper names, general academic words and domain-specific words. As in the English Academic Spoken Word list (Dang et al., 2017), we will divide the list into sublists of 50 words, based on their frequency. Additionally we will process the corpus with term extraction tools such as D-terminer (Rigouts Terryn et al., 2022), TermTreffer[9] or their successors.

We will also develop a frequency-based spoken academic vocabulary test targeting students' aural recognition of academic word forms, as well as their meanings. The test will be divided into test sections, corresponding to the sublists of the frequency list. It will also have an online multiple choice format and students will be provided with the spoken form of the word and will have to tick off the correct option which corresponds to the word's meaning. The first test version will be piloted with a small group of Dutch-speaking students (n=30) before the start of the actual larger scale validation process.

For the moment, exploring and implementing further practical applications, as well as evaluation of the corpus are still areas of future research. Of course, once the corpus has been made available to researchers, a multitude of other uses and applications can be envisaged too, such as comparisons at lexical, syntactic and other levels with other (spoken and/or written) Dutch corpora.

Acknowledgments

The SABeD project is funded by KU Leuven Internal Funding, Research Project 3H200610, with additional support from the Instituut voor de Nederlandse Taal.

Use of CLARIN-NL tools on the SABeD data was work done in ClaSABeD (Clariah-NL tools in SABeD), funded by CLARIAH-PLUS Work package 3, project no. CP-WP3-22-001.

We would like to explicitly thank the transcribers of the corpus for their dedicated and tedious work: Israe Ahale, Ellen Boucké, Mirne De Beule, Gijs Heynickx, Emma Janssens, Yitse Murillo, Kim van den Berg and Joke Van Mol.

References

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*, 243–257.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, *36*, 9–48.

[9] https://termtreffer.org/

Bonne, P., & Casteleyn, J. (2022). Taalbeleid en taalondersteuning: Op zoek naar een gedeelde basis en strategie voor implementatie. *Tijdschrift voor Onderwijsrecht en Onderwijsbeleid*, *4*, 279–293.

Dang, T. N. Y. (2022). Vocabulary in academic lectures. *Journal of English for Academic Purposes*, *58*, 101–123.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67(4)*, 959–997.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2021). Vocabulary in academic spoken English. *New Zealand Studies in Applied Linguistics*, *26(2)*.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating Research Environments with BlackLab. In *CLARIN in the Low Countries*. Ubiquity Press.

Deygers, B. (2017). Validating university entrance policy assumptions. Some inconvenient facts. In E. Gutiérrez Eugenio (Ed.), *Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference* (pp. 46–50). Cambridge: ALTE.

Deygers, B., & Malone, M. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36(3)*, 347–368.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, *32*, 43–56.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing*, *35*, 449–476.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*.

Heeren, J., Speelman, D., & De Wachter, L. (2020). A practical academic reading and vocabulary screening test as a predictor of achievement in first-year university students: Implications for test purpose and use. *International Journal of Bilingual Education and Bilingualism*, *0(0)*, 1–16.

Heeren, J., Speelman, D., & De Wachter, L. (2021). Bepaalt taal wie het haalt? de samenhang tussen een academische taalvaardigheidscreening en het behalen van een bachelordiploma aan de universiteit. *Tijdschrift voor Hoger Onderwijs*, *39(1)*, 39–54.

Kessens, J. M., & van Leeuwen, D. A. (2007). N-best: the northern- and southern-Dutch benchmark evaluation of speech recognition technology. *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 1354–1357. http://www.isca-speech.org/archive/interspeech%5C_2007/i07%5C_1354.html

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistic Review*, *4(1)*, 151–172.

Nation, I. S. P. (2016). Making and using word lists for language learning and testing.

Oostdijk, N., Goedertier, W., van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002, May). Experiences from the spoken Dutch corpus project. In M. González Rodríguez & C. P. Suarez Araujo (Eds.), *Proceedings of the third international conference on language resources and evaluation (LREC'02)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/98.pdf

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit [IEEE Catalog No.: CFP11SRW-USB]. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2022). D-terminer : online demo for monolingual and bilingual automatic term extraction. In R. Costa, S. Carvalho, A. Ostroski Anic, & A. F. Khan (Eds.), *Proceedings of the Workshop on Terminology in the 21st century : many faces, many places* (pp. 33–40). European Language Resources Association (ELRA). %7Bhttps://lt3.ugent.be/dterminer/%7D

Simpson, R. C., Briggs, S. L., J., O., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31(4)*, 577–607.

Szudarski, P. (2017). *Corpus linguistics for vocabulary. A guide for research*. Routledge.

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, *5*, 263–264.

Tiberius, C., & Schoonheim, T. (2013). *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge.

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, *22*, 349–365.

Uchihara, T., & Webb, S. (2022). Materials for teaching vocabulary. In *The routledge handbook of materials development for language teaching* (pp. 202–217). Taylor; Francis.

van den Bosch, A., Busser, G., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). Centre for Computational Linguistics.

Van Dyck, B., BabaAli, B., & Van Compernolle, D. (2021). A Hybrid ASR System for Southern Dutch. *Computational Linguistics in the Netherlands Journal*, *11*, 27–34. https://clinjournal.org/clinj/article/view/119

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied linguistics*, *34(4)*, 457–479.

Vandeghinste, V., & Guhr, O. (2023). FullStop: Punctuation and Segmentation Prediction for Dutch with Transformers. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09676-x

van Gompel, M., & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, *3*, 63–81. https://clinjournal.org/clinj/article/view/26

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72(1)*, 5–44.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1556–1559.