JS Journal of HistoricalSyntax
Volume 6, Article 6: 1–36, 2022

CHALLENGES IN TAGGING AND PARSING SPOKEN DIALECTS OF DUTCH*

Melissa Farasyn
Ghent University

Anne-Sophie Ghyselen
Ghent University

Jacques Van Keymeulen
Ghent University

Anne Breitbarth
Ghent University

Abstract This paper reports on the construction of a tagged and parsed corpus of the southern Dutch dialects. The corpus aims to facilitate diachronic research into the syntax of Dutch, as its dialects have retained many interesting (morpho)syntactic features, which can often be traced back to changes starting in or characteristics retained from older stages of historical Dutch. The discussion mainly focuses on initial test results achieved by applying existing NLP tools, which have been developed or optimized for POS tagging and parsing modern standard Dutch. We report on initial tests on our data with Frog, TreeTagger and Alpino. We discuss some of the challenges we have encountered working with spoken, unstandardized language in general on the one hand and on specific (morpho)syntactic problems for POS tagging and parsing the southern Dutch dialects on the other hand. The challenges and solutions we present in this pilot study will inform our choices

* The *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* is currently funded by a medium-size research infrastructure grant from the *Fonds voor Wetenschappelijk onderzoek – Vlaanderen* (*FWO*) (grant number I010120N, 2020–2024) and a postdoctoral fellowship from FWO to Melissa Farasyn (grant number 12P7919N, 2018–2021). During the pilot phase (2018–2019), it was furthermore funded by a small research grant from FWO to Anne Breitbarth (grant number 1.5.310.18N) and by grants from the provinces of West Flanders (KIOSK-projectID 29575), East Flanders (KIOSK-projectID 29311) and Zeeland (reference of confirmation letter: 19004695) to Variaties vzw. We are indebted to all the volunteers and student annotators, as well as, in particular, Lien Hellebaut, who supervized and guided them. We furthermore thank Prof. Gertjan van Noord of the University of Groningen for running the first experiments with the Alpino parser on our data.

for the NLP tools we will use or adapt for the development of the fully annotated corpus.

1 Introduction

Studying dialects and other non-standard language varieties can contribute significantly to refining and enriching theoretical models of language structure in syntax. For this reason, dialects have gained a lot of attention in the literature on syntax during the last decades (e.g. Barbiers, Koeneman, Lekakou & van der Ham (eds.) 2008b), and have lead to a number of relevant infrastructure projects, e.g. on Northern Italian dialects (Poletto & Benincà 2007), Scandinavian dialects Lindstad, Nøklestad, Johannessen & Vangsnes (2009), various German dialect areas (e.g. the projects presented in Kehrein, Lameli & (eds.) 2015, but also Herrgen 2010, Brandner 2015, Fleischer, Lenz & Weiß 2017), and also Dutch (Barbiers & Bennis 2007, Barbiers, Bennis, Vogelaer, Devos & van der Ham 2005, Barbiers, Bennis, Vogelaer, van der Auwera & van der Ham 2008a).[1] Additionally, studies in dialect geography contribute to research in historical language change, since microvariation in and diachronic change of a given structure are inextricably connected with each other. This is exactly what the adage of the Junggrammatiker, *Aus dem räumlichen Nebeneinander ein zeitliches Nacheinander*, reflects: the linguistic landscape at a certain point in time reflect chronological stages of a language in different regions in space.

An excellent example of a linguistic landscape that reflects diachronic language change is the Dutch language area, whose dialects have retained many interesting (morpho)syntactic and other linguistic features from older stages of Dutch. In the southern Dutch dialects, such properties are well-preserved due to later dialect loss (as a result of increasing influence of the standard language through schools and media as well as increased mobility) compared to many other European language areas, and to the very low average literacy rate of the Flemish population until the 20th century.

This article describes the *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* (GCND) ('Spoken corpus of the southern Dutch dialects') and its construction. The focus in the current article will be on the initial test results yielded by experiments on annotating and parsing transcriptions of recordings of spontaneous speech in southern Dutch dialects with existing NLP tools, which were originally designed for the linguistic annotation of Standard Dutch corpora. Thanks to the development of tagged and parsed diachronic corpora, more and more quantitative studies of syntactic changes

1 For an overview of more projects, see http://www.dialectsyntax.org/wiki/Welcome.

can be performed. This is because linguistically-enriched corpora facilitate searching large amounts of data and deliver easily reproducible results. We motivate the choice of the tools we will use for the development of the GCND and the enrichment of the Dutch dialect data on the basis of the observations we made in the experiments described in this article.

1.1 Southern Dutch dialects in historical linguistics

With the term 'southern Dutch dialects' (SDDs), we refer to the dialects spoken in (i) Dutch-speaking Belgium, (ii) the three southern provinces of the Netherlands (Limburg, Noord-Brabant and Zeeland) and (iii) the Flemish-speaking dialect region in France. In this area, four larger dialect groups can be distinguished (see Figure 1): The first one is Flemish, further subdivided into (a) West-Flemish, (b) East-Flemish, (c) Zeeland-Flemish, spoken in the southern part of the Dutch province Zeeland Flanders, and (d) French-Flemish, the nearly extinct West-Flemish dialect spoken in the northwestern part of Nord-Pas-de-Calais in France. The second dialect group is Zeelandic, spoken on the Zeelandic islands and the South-Hollandic islands Goeree and Overflakkee. Third, Brabantic dialects are spoken in the provinces North-Brabant, Antwerp and Flemish Brabant. The last dialect group is Limburgish, spoken in Belgian and Dutch Limburg.

The southern Dutch language area is, historically, a politically fragmented area, in which official documents as well as literary texts were written in Latin as well as in different dialects (Marynissen & Janssens 2013). In the 13th and 14th century, it was especially the western County of Flanders which flourished economically and culturally. This is reflected in the amount of Flemish texts which were produced in this period. In the Corpus Gysseling, for instance, which collects all preserved Dutch texts from before 1301, approximately three-quarters is Flemish. The economical and cultural centre later shifted from Flanders to Brabant in the 14th and 15th century, which lead to an increasing number and variety of texts written in the Brabantic dialects (Marynissen & Janssens 2013). The first attempts to standardize the language, based on the Brabantic dialects, took place in the southern Dutch language area. In the 16th century, for example, the first grammars and spelling guides were printed. After the fall of Antwerp in 1585, the centre of standardization shifted to the province of Holland. Codification and stardardization evolved further in the protestant north during the 17th century (Van der Wal 1995). Meanwhile, in the catholic south, French was the dominating language in the highest levels of administration, economy, culture and education from the 17th century onward, except for a short period under the rule of Willem I of the Netherlands from 1815 until 1830. It was only in the second half of the

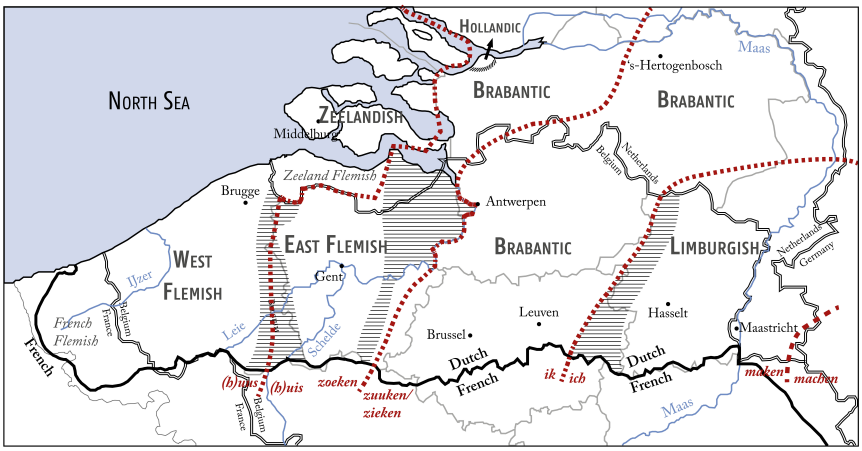

Figure 1 Southern Dutch dialects and transition zones (based on Taeldeman (2001), transition zones between dialect groups are hatched)

19th century that cultural and linguistic rights for Dutch speakers were explicitly fought for, leading to the recognition of Dutch as an official language in Belgium, next to French, in 1898 (Marynissen & Janssens 2013). In colloquial speech, however, people had continued to use their native language instead of French for centuries. Because of the late official recognition of Dutch in Flanders, the need for a standard language was only felt quite late, compared to other European speech communities. As such, dialect levelling and shift, a typical side-effect of standardization, also set in quite late in Flanders (Vandekerckhove 2009, Ghyselen & Van Keymeulen 2014). The late introduction of general compulsory education (not until 1914) and the very low average literacy rate until then also played a major role in the preservation of the dialects.

The (language) situation was very different in the Flemish speaking language area in northern France, which is usually referred to as French Flanders. That area was originally part of the County of Flanders, in which, historically, the Flemish dialect used to be the dominating language, whereas French was only used occasionally to address French-speaking lords (Ryckeboer 2013).

From 1678 onward, after the signing of the Peace Treaty of Nijmegen, the region was no longer part of the County of Flanders and Flemish gradually lost its dominance in the region. The regional Flemish, often referred to as French Flemish, is now moribund. The remaining speakers are born before World War II and are all bilingual. There is no native language acquisition of the Flemish dialects anymore (Ryckeboer 2013). Because of its western position and the 340 years of isolation from the other SDDs and later Standard Dutch, French Flemish is quite distinct, typologically, from the other SDDs and even from Belgian West Flemish dialects that are most closely related to it.

The southern Dutch language history shows that diachronic research into Dutch as a whole is impossible without taking into account its dialects, as they form a missing link in the language history since Middle Dutch (Marynissen & Janssens 2013). The SDDs form an especially interesting research topic, as they display numerous typological peculiarities and differ in many aspects both from the overarching standard language, which they are related to, and from each other. Some of these syntactic aspects, which are not found in the overarching standard language, are for instance the retention of the old negation particle *en* in certain contexts, whereas there is at the same time exaptation of the element (see, among others, Overdiep 1933, Haegeman et al. 1995, Neuckermans 2008, Breitbarth & Haegeman 2014), verb later than second (V>2) constructions (a.o. Haegeman & Greco 2018), subject doubling after a complementizer (De Vogelaer & Devos 2008), complementizer agreement (Haegeman & Van Koppen 2012), subject cliticization after *ja* ('yes') and *nee* ('no') (see, among others, Haegeman 1992, Barbiers et al. 2008a) and a large variety of discourse particles (Haegeman & Hill 2013).

1.2 Voices from the past

Many of the unique characteristics of the SDDs only occur in very specific discourse contexts that are difficult to elicit in constructed experimental settings such as questionnaires. Existing dialect collections for syntactic research are, however, often based on elicited data and are therefore not always sufficient to fully map or to even notice certain phenomena.[2] On top of that, the use of elicited data of contemporary dialect speakers is only a partial solution, especially considering the now advanced dialect loss in Flanders (Vandekerckhove 2009, Ghyselen & Van Keymeulen 2014). Therefore, elicited data should

2 Haegeman & Greco (2018: 8–9) for instance point out that there can be a clear discrepancy (apparent from the SAND field notes on the DynaSAND website) between acceptance and actual production in certain syntactic patterns in the SAND, in their case non-inverted subject-initial verb second after clause-initial central adverbials (SAND sentence 359, DynaSAND, Barbiers et al. 2005: 74, map 95a), cf. also Section 4.3.1 below.

ideally be complemented by spontaneously spoken dialect data by speakers who acquired the local dialect as their first, and preferably only, language (cf. already Blancquaert 1948: 12).

Section 1.1 described how the French Flemish dialects were losing their dominance in French Flanders. In the 1960s, the French Flemish dialect already started to go extinct, as there was no first-language acquisition anymore. The other SDDs, which were not roofed by the French standard language, were retained reasonably well up to the 1970s, until dialect loss gradually started to occur there as well. In the beginning of the 1960s, two dialectologists from Ghent University, Prof. Willem Pée and Prof. Valère Vanacker, recognized the need to capture the SDDs before they were lost. Consequently, between 1961 and 1978, they made (or commissioned) recordings in about 550 different locations in the whole southern Dutch language area (Vanacker & De Schutter 1967). The regional distribution of these recordings is shown in Figure 2. The recordings contain spontaneously spoken dialects in their natural form, spoken by what are generally considered reliable dialect speakers, i.e. speakers that are non-mobile, of a certain age (born around 1900), rural, unschooled, preferably with parents who grew up in the same locality as well (cf. the NORM-speaker, Chambers & Trudgill 1998). Because the speakers were born around 1900 (the oldest in 1871), the corpus also represents a historical language stage: the recordings provide an insight into the traditional dialects spoken in the southern Dutch language area in the first half of the twentieth century. The original goal of the recordings was to document the local dialects, but also to facilitate the study of their dialectical, and especially syntactic, characteristics (Vanacker & De Schutter 1967). The enormous collection contains 783 recordings (approximately 700 hours) of free conversations in the local dialects of speakers in 550 different locations.

The recordings were originally captured on reel-to-reel tape, but were digitized in 2014. This digital collection, labelled *Stemmen uit het verleden* ('Voices from the past') is freely available online via the website of Dialectloket.[3] The collection of recordings is currently only searchable for keywords relating to their content, but not for any linguistic features.[4] These keywords enable historical research into the content of the recordings in particular and on oral history in general.

3 http://www.dialectloket.be/geluid/stemmen-uit-het-verleden/

4 Thanks to funding by a project grant for international cultural heritage of the Flemish department of youth, culture and media, co-financed by Variaties vzw (grant 037087), these keywords were recently standardized by means of a thesaurus.

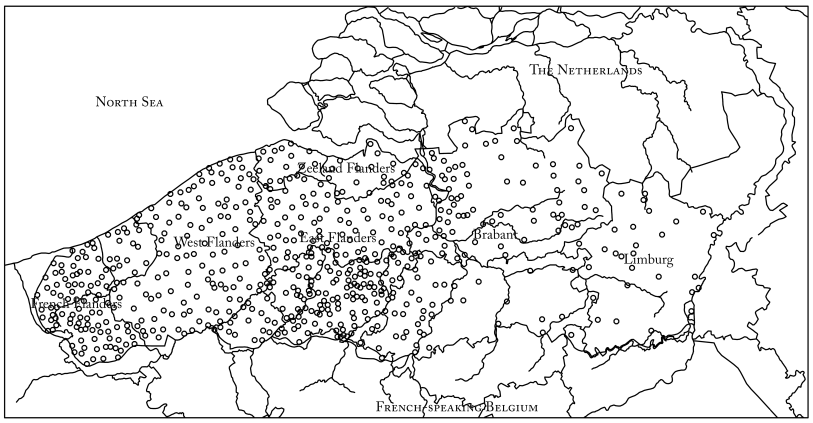

Figure 2 Regional distribution of the recordings.

When the dialect recordings were made, a number of recordings were transcribed (318 in total), sometimes by students in the context of a licentiate's dissertation, by student assistants, or by volunteers. However, the available transcriptions are not optimal for further research. On the one hand, the large quality differences between the transcriptions are problematic: some of them are typed and are easy to read, in others the ink has faded or there are remarks and corrections between the lines and in the margins. After all, each transcription was first written out by hand, then checked and finally typed out. Some transcriptions, however, did not reach that final stage, or the final typescript was lost over the years, so that a large number of the transcriptions is only preserved in hand-written form. As a result, the OCR of most of the collection would be labour-intensive, requiring a great deal of manual correction. An even bigger problem is the lack of a uniform transcription protocol. At the time, only very brief transcription guidelines were available, resulting in a large amount of variation in the way in which dialect characteristics are represented orthographically.

1.3 The GCND

The GCND is the first linguistically-annotated corpus of spoken Dutch dialects, bridging the gap between Middle Dutch, Early Modern Dutch, and

modern Dutch language resources. The project aims to make the dialect recordings described in Section 1.2 accessible for linguistic research. Compared to other data collections on Dutch dialects, and indeed other historical corpora, the GCND is unique in being based exclusively on spontaneous speech. So far, 351 dialect recordings covering the three large dialect groups within the SDDs have been uniformly orthographically transcribed using the transcription software ELAN,[5] covering about 45% of the entire collection, and 64% of all locations, as there is more than one recording for some locations.

The transcription in ELAN makes it possible to align multiple layers of markup immediately with the audio. The transcriptions are linguistically annotated with Part-of-Speech (POS) tags and parsed, i.e. the syntactic functions and relations of word groups and clauses are allocated and visualized. In light of the CLARIN philosophy, which encourages and supports the sharing, reusing and sustainability of research tools in the humanities (cf. de Jong, Maegaard, de Smedt, Fišer & Uytvanck 2018), we rely on existing natural language processing (NLP) tools and tagsets as much as possible. As the dialect recordings represent a historical stage of the language (in the case of French-Flemish even the last witness of a now all-but-extinct language variety) and will be searchable for word forms and syntactic patterns, the GCND will (i) make it possible to track language change through time and space, (ii) enable a new perspective on the functional strength of dialect features in real life and (iii) facilitate the serendipitous discovery of previously unnoticed structures. Although the transcription itself is not phonetic, time-alignment between audio and transcription facilitates phonetic research. Audio, audio-aligned transcriptions and annotations will be made available online with query tools.

2 Considerations on the linguistic enrichment of spoken dialect data

NLP tools are typically trained on contemporary standard languages. In recent years, more and more research has been conducted into the adaptation of existing NLP tools for low-resource languages (e.g. Zampieri, Nakov, Malmasi, Ljubešić, Tiedemann & Ali 2019). For these languages, contrary to high resource languages such as most Western European standard languages, parallel data resources such as dictionaries and grammars usually are rare or not even available at all. The data in the pilot project of the GCND are challenging to automatically enrich linguistically in at least two ways. In the first place, the corpus contains highly variable dialect data, which can be considered low-

5 https://archive.mpi.nl/tla/elan. For further details on the transcription protocol and the decisions leading up to its establishment, cf. Ghyselen, Breitbarth, Farasyn, Van Keymeulen & van Hessen (2020a), Ghyselen, Keymeulen, Farasyn, Hellebaut & Breitbarth (2020b).

resource languages. Notwithstanding the often-made division of the Southern Dutch dialects into four large groups (Flemish, Zeelandic, Brabantic and Limburgian), the structural variation is considerable, even within groups. It is often the case that even the dialects of neighbouring villages differ from one another on a phonological, lexical, morphological or even syntactic level. Secondly, we are dealing with transcriptions of spoken languages instead of with originally written texts, the latter being the standard in historical linguistics. This poses some interesting challenges for existing NLP tools for linguistic markup such as POS taggers and parsers, as they are usually made for and trained on written texts, which usually contain punctuation and show less hesitations, reformulations or unfinished sentences.

This Section describes some of the initial challenges we have encountered due to the fact that our data display a wide range of variation and are based on spoken data. We describe some decisions that were made to address these issues in the transcription phase already instead of in the POS and parsing phase. For a full description of the transcription protocol, we refer to Ghyselen et al. (2020a); in this Section, we only address the decisions which are of importance to make the later enrichment easier. The most important decision in that respect is that the transcriptions of the speech of the dialect informant always consist of two layers, one closer to the dialect, and one closer to Standard Dutch. The layer closer to Standard Dutch is the one to which the POS tagger and the parser are eventually applied.

In the first transcription layer, non-standard language vocabulary, morphology and syntax are preserved and transcribed according to Dutch spelling rules. Most of the phonological variation is standardized however, as a uniform and accurate representation of this variation would require IPA (or another phonetic alphabet). Such IPA transcription would make the transcription procedure too complicated, time-consuming and expensive (cf. Ghyselen et al. 2020a). However, the alignment with the audio ensures that the sound forms remain accessible. The only phonetic/phonological non-standard language features marked in the first layer are deletions and insertions of consonants in function words (for example, we transcribe *me* instead of *met*), as these features can be fairly unambiguously transcribed without phonetic characters. The general standardization of phonetic/phonological variation makes it easier to classify words into different part-of-speech categories automatically, as the variation in the data set (and the amount of training data needed) gets smaller. Clitic elements – such as *k* in *k#weten* ('I know') – occur frequently in the spoken dialects. In the first transcription layer, they are transcribed as clusters, with the individual parts of the cluster separated with a #. Marking the boundaries between the clitic elements facilitates alignment

with the second layer, in which the elements are separated into different tokens which will each receive their own POS tag.

In the second layer, not only phonology, but also morphology is standardized. *Weten* in *k#weten* ('I know') for instance, a first-person singular inflected verb which carries the non-standard suffix *-en*, becomes *weet*. This makes it easier for the end-user to find all possible occurrences and ways in which a certain morphological form may occur without extensive knowledge of the particular dialect. Furthermore, it ensures that the POS tagger recognizes the particular form more easily. Non-standard language syntax and vocabulary, however, are preserved in the second layer, as it is often difficult to 'translate' those lexical and syntactic attributes into Standard Dutch. For these elements, the second layer is therefore a standardization, not a translation. The second layer allows users without extensive knowledge of the dialect to search the data.

Although spoken language does not contain punctuation, we have decided to let the transcriber manually insert limited punctuation, consisting of full stops, question marks and ellipsis (...). They indicate where a sentence is interrupted or ends. Full stops (and question marks) are solely used for that purpose, and thus not combined with letters for abbreviations (i.e. not 'A.B.N.' but 'ABN') or numbers (i.e. not '1,000', but 'thousand'). Unlike many historical languages that are only transmitted through written sources, deciding on this limited punctuation is possible on the basis of intonation, pauses and dialect knowledge of the transcribers. Adding punctuation marks is particularly important for parsers, which usually partly rely on punctuation marks in order to identify a meaningfully informative coherent sentence or phrase and find the relevant constituents in it.

Example (1) illustrates the transcription procedure in two layers with a passage from the transcription of Broekkerke (French Flanders):

(1)

a	*k#en#en#n#ik*	*noois*	*willen*	*leren.*
ah	ik en heb hem ik	nooit	willen	leren.
ITJ	I NEG have him I	never	wanted	learn

''Ah I have never wanted to learn it (= the catechism).''

(S007p Broekkerke)

As can be seen from this example, the vowels in the first layer are transcribed according to standard spelling; no attempt is made to record the dialectical pronunciation that can be heard in the recording orthographically. The dialectical function words (such as the pronoun *k* for *ik*, 'I'), clitic clusters (k#en#en#n#ik) and inflectional morphology (*en* for *he*(*bbe*)*n*, 'have'), on the other hand, are represented approximating the dialectal form. In the second

layer the clitic cluster is resolved and the function words (here *ik*) are written according to standard conventions, while the dialectical word order is maintained ('ik NEG heb hem ik' vs. Standard Dutch 'ik heb hem NEG'). Example (2) from Hondegem illustrates some of the many interjections typically used in spoken dialects (*eum*, *eneeë*). These are, as well as the dialect lexicon, not translated in the second layer, since a close discourse study of the distribution of these particles would be needed to know their exact meaning.

(2)

eum	*je*	*passeert*	*en*	*je*	*zijt*	*bij*	*die*	*ofsteden*	*eneeë*
eum	je	passeert	en	je	bent	bij	die	hofsteden	eneeë
ITJ	you	pass by	and	you	are	with	those	homesteads	ITJ

''You pass by and you arrive at those homesteads.''

(N108p Hondegem)

In Section 3 we evaluate the application of two existing POS taggers. In Section 4 we discuss the use of parsers for (varieties of) modern Dutch on transcriptions of dialect tape recordings of the SDDs. We also discuss some issues which could not be solved in the transcription phase yet.

3 Part-of-Speech tagging

A POS tagger automatically classifies tokens into categories of words. The accuracy of the tagger depends among other things on the amount of manually labelled data, i.e. a so-called gold standard, on which it has been trained, but also on the variation in the data. If the training data did not comprise a lot of language variation and the dataset which needs to be tagged does, the tagging accuracy could be very low. When trained on more heterogeneous input, however, the accuracy of automatically assigned tags can be very high (though more training data will be needed, too, as high variation leads to sparsity). For example: the POS tagger of the Corpus of Historical Low German (CHLG), which was trained with supervized learning (i.e. using preset parameters), reaches a Global Accuracy of 87.7% (Koleva, Farasyn, Desmet, Breitbarth & Hoste 2017: 135), while the customized tagger for the Middle Welsh corpus reaches a Global Accuracy of 90.4% (Meelen 2020). Building such a classifier from scratch for dialectological data from the SDDs would, however, require the manual labeling of huge amounts of data because of the immense linguistic variation in the dataset.

An important characteristic of POS taggers which should be kept in mind when evaluating existing software, is that every POS tagger makes use of a particular tagset, which can be more or less detailed, depending on the lan-

guage and the needs of the corpus. This issue will be dealt with more elaborately in Section 3.1.

3.1 Data selection

Section 1.1 described how the dialectal landscape in the southern Dutch language area shows much variation on domains represented in the transcriptions such as lexicon, morphology and syntax. Even geographically very close dialects might thus differ greatly on several structural levels of the language. Consequently, when making or using natural language processing tools, it is important to find a robust classifier which obtains a high classifying accuracy, even when there is considerable variation in the dataset.

For the present paper, we tested two commonly used taggers for Dutch, TreeTagger and Frog, on a representative selection of transcriptions from the GCND corpus under construction. Both were applied to the second transcription layer, which is closer to Standard Dutch. These two taggers were chosen mainly for their differences in tagset and training data. TreeTagger offers a rather small tagset for Dutch, containing 42 tags, while the POS tagger of Frog for Dutch consists of about 300 tags. The selected data consist in one transcription for each dialect area and and each transition zone (between dialect areas): in this paper we focus on the recordings made in Hardifort (French-Flanders), Ypres (western West-Flanders), Oudenburg (coastal West-Flanders), Gent (East-Flanders), Wesdorpe (Zeeland-Flanders), Sint-Joris-Weert (Brabantic), Uikhoven (Limburgish), Maldegem (transition zone between West- and East-Flanders), Sint-Niklaas (transition zone between East-Flanders and Brabantian). Currently, there is no recording from the dialectal transition area between Brabantian and Limburgish in the test corpus.[6] The chosen places are illustrated in Figure 3.

3.2 Existing toolkits and classifiers for POS tagging Dutch

The first tagger that we have tested on our data is TreeTagger, which was developed at the Institute for Computational Linguistics of the University of Stuttgart (Schmid 1994, 1995).It is an NLP tool for annotating texts, which provides POS tags, lemma information and chunking. The tool is language-

6 With the medium-sized infrastructure funding from the FWO (2020–2024; I010120N), several new recordings will be made to fill the gaps in the collection towards the East (cf. also Fig. 2). Due to the ongoing Corona pandemic, which makes the recording of elderly speakers rather difficult or outright impossible, this part of the project had to be delayed for the moment. As no transcriptions had been made yet of Zeelandic recordings at the time the analysis, no Zeelandic data were included included in the dataset for this paper.

Figure 3 Transcriptions considered in the pilot study of the tagger for POS tagging and parsing.

independent, as it can be used on or adapted for any language as long as a lexicon and manually tagged training data are available. The software is freely available for further research, education and evaluation, which means that it is also adaptable to other languages whenever a lexicon and a gold standard training corpus are provided. TreeTagger has been used successfully to tag Dutch data already and as a consequence, a parameter file including a Dutch language model is already available online.[7] The parameter file and the POS tagset we tested are freely available for research on the website of TreeTagger. The Dutch Web Corpus (NlWaC) POS tagset was developed for tagging Dutch.[8] The parameter file has been developed in order to tag the Dutch Web Corpus (nlTenTen) with TreeTagger, which belongs to the TenTen Corpus Family.[9]

The other NLP toolkit we have tested is Frog. Frog is a set of NLP tools for Dutch based on the memory-based learning software package TiMBL (Daelemans, Zavrel, Van Der Sloot & Van den Bosch 2004), the modules of which were mainly developed at the ILK Research Group at Tilburg University and

7 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
8 https://www.sketchengine.eu/dutch-nlwac-tagset/
9 https://www.sketchengine.eu/nltenten-dutch-corpus/

the CliPS Research Centre of the University of Antwerp (Sloot, Hendrickx, van Gompel, den Bosch & Daelemans 2018). The different modules developed over the years are now integrated into one tool, which is currently maintained and developed by the Language Machines Research Group and the Centre for Language and Speech Technology at the Radboud University in Nijmegen. With Frog, plain text files can be tokenized, POS tagged, lemmatized and parsed. It also offers morphological segmentation of the lexemes. Frog produces FoLiA XML or tab separated files as output and is run on plain text input. The software is available under a GNU General Public License and can be used and modified for further research.

The POS tags assigned by Frog are based on the tagset of the *Corpus Gesproken Nederlands* ('Corpus of Spoken Dutch', CGN). It consists of 12 general tags and 280 more fine-grained tags.The POS tagger is trained on a Dutch corpus of 10,975,324 tokens, 90% of which comes from the transcribed CGN. A confidence score, which indicates how certain the classifier is of the assigned tag, is assigned to each POS tag as well.

3.3 Results

We automatically POS tagged a transcription from each of the nine areas described in Section 3.1 with the two POS taggers discussed above. To evaluate their performance, the first 1000 tags from each transcription were manually-corrected by a linguist with a profound knowledge of Dutch, i.e. we have 18,000 manually corrected POS tags, based on the tokens of the transcription layer which is closer to Dutch.

3.3.1 TreeTagger

The POS tags provided in the Dutch parameter file for TreeTagger are specifically developed for this tagger. The tagset includes subdivisions for singular and plural verbs and nouns, tenses, and types of pronouns, adverbs and conjunctions, but are not complex on a further morphological level, for instance case or number. The fact that the tagset is rather small helps to ensure a high accuracy. As Table 1 shows, it reaches an average accuracy of 92.9%, with a maximum of 96.3% in the dialect of Sint-Niklaas, which is located in the transition area between the East-Flemish and the Brabantic area, whereas the accuracy reaches its lowest point (90.1%) when tagging the archaic French Flemish dialect of Hardifort.

Place	**Correct**	**Incorrect**	%	
Oudenburg (H24)	942	58	94.2	1000
Maldegem (I154)	947	53	94.7	1000
Westdorpe (I166)	909	91	90.9	1000
Sint-Niklaas (I175)	963	37	96.3	1000
Gent (I241)	918	82	91.8	1000
Ypres (N72)	904	96	90.4	1000
Hardifort (N94)	901	99	90.1	1000
Sint-Joris-Weert (p130)	918	82	91.8	1000
Uikhoven (Q013)	959	41	95.9	1000
Total	8361	639	92.9	9000

Table 1 Accuracy of POS tagging GCND transcriptions with TreeTagger

Wrongly classified tokens include (i) unrecognized prepositions (example 3 a),[10] which are not or differently used in Standard Dutch, (ii) various interjections such as *awel* 'well', *eni* 'right, isn't it', or *zenne* 'you see' (example 3 b), (iii) proper names (example 3 c) and (iv) dialectal adverbs which do not occur in contemporary spoken Standard Dutch such as *algelijk* 'still, whatsoever, anyway', *stijf* 'very' and *altemets* 'sometimes' (example 3 d).

(3) a. *mijn vader de man euh... hij wrocht altijd* ***naar*** *Paul Quidts.*
my father that man ITJ he worked always at Paul Quidts
'My father always worked at Paul Quidts' place.' (N94 Hardifort)

b. *dat is een antiquiteit* ***zenne***
that is an antiquity ITJ
'That is an antiquity.' (I175 Sint-Niklaas)

c. *en hoe is het bij jullie op de* ***Vrijdagsmarkt?***
and how is it with you on the Vrijdagsmarkt
'And how are you doing on the Vrijdagsmarkt?' (I241 Gent)

d. *ik vind dat dat* ***stijf*** *goed is*
I find that that very good is
'I think that that is very good.' (H24 Oudenburg)

10 In the examples in this Section we use the second transcription layer (closer to Dutch), as it is the layer the POS tagger was applied to.

Furthermore, in certain aspects, the tagset is not detailed enough. For example, it contains no specific tag for dialectal negation particles such as *en* as can be seen in example (4), which are - if recognized at all - more generally tagged as adverbs; for the specific problems concerning *en* cf. Section 4.3.2 below. The same problem arises with the verbal particles of separable verbs, which can also bee seen in example (4), where the particle *op* 'up' is separated from the rest of the verb *letten* 'pay attention' by the infinitival marker *te* 'to'. Similar examples are common in the test corpus. The tagset does however contain the tag 'partte' for the infinitival marker *te*, here between *op* and *letten*.

(4) *en als je het verstand niet* ***en*** *had van een beetje*
and if you the intellect NEG NEG had of a little
op te letten *je ging al je alaam je harnas rampeneren*
to pay attention you went all your stuff your frame destroy
''And if you would not have the intellect to pay a little attention, you would break the whole frame.''

(N94 Hardifort)

Finally, some unrecognized items and wrongly assigned tags can be expected. This is for instance the case for the token *ggg*, which we use for indicating fragments in which the informant coughs or laughs (cf. Ghyselen et al. 2020a), as shown in example (5), which was not part of the training data for TreeTagger.

(5) *ze ging er een keer een man van maken zei ze. ze ging*
she went ER a time a man of make said she she went
hem leren ***ggg***.
him learn ggg
'She was going to make a man out of him she said. She was going to teach him.'

(I241 Gent)

3.3.2 Frog

Table 2 shows that Frog generally reaches the higher accuracy of the two classifiers, with an average accuracy of 94.5% for the dataset. The best result, 98.8%, is obtained in the Flemish dialect of Westdorpe, which is located in Zeelandic Flanders. The lowest accuracy results on the other hand occur when tagging the Brabantic dialect in Sint-Joris-Weert (89.3%).

The largest group of tokens which are tagged incorrectly are different kinds of interjections, such as *awel*, *hé* or *zenne*. These are usually tagged as

Place	**Correct**	**Incorrect**	%	
Oudenburg (H24)	952	48	95.2	1000
Maldegem (I154)	973	27	97.3	1000
Westdorpe (I166)	988	12	98.8	1000
Sint-Niklaas (I175)	946	54	94.6	1000
Gent (I241)	929	71	92.9	1000
Ypres (N72)	926	74	92.6	1000
Hardifort (N94)	938	62	93.8	1000
Sint-Joris-Weert (p130)	893	107	89.2	1000
Uikhoven (Q013)	959	41	95.9	1000
Total	8504	496	94.5	9000

Table 2 Results on accuracy of POS tagging GCND transcriptions with Frog

nouns or adjectives instead, as was the case with TreeTagger (see example 3 b). In the transcription of Sint-Niklaas for instance, all the incorrectly tagged tokens are such interjections. Furthermore, the preverbal negative particle *en* is usually tagged as the conjunction *en* 'and', which is spelled the same (see example 4). It is interesting to note that *nee* ('no'), used to give a negative answer, is always tagged correctly as an interjection (ITJ), whereas the variant *neen* ('no'), with the same function, is consistently classified as a noun (N) instead. *Nee*(*n*) can be used as a noun as well (for instance in the phrase *Een nee*(*n*) *heb je, een ja kun je krijgen* 'You have a no, you can get a yes'), but it does not appear in our data as such.

(6) ***neen neen*** *dat was zo erg niet nee*
no no that was so bad not no
''No, it was not that bad, no.''

(P130 Sint-Joris-Weert)

A positive aspect of Frog is that it does recognize some dialect lexemes, and tags them correctly, e.g. *bè* as a dialectal interjection `TSW(dial)`, cf. (7). This is probably because these non-standard features also sometimes occur in intended Standard Dutch, and the spoken CGN data on which the POS tagger was trained also contain these elements.

(7) ***a*** *ja* ***bè*** *we hebben wij gewoond hier langs de vestingen*
ITJ yes ITJ we have we lived here along the strongholds
in een barak
in a barrack
''Yes, we lived here along the strongholds in a barrack.''

(N72 Ypres)

As was the case for TreeTagger, Frog often misclassifies the token *en*. Its most common use as a conjunction ('and'), as it is used in Standard Dutch as well. However, it can, as mentioned above, also be a preverbal negation particle in the dialects, which is, as was the case with TreeTagger (see example 4), usually not recognized.[11] A last observation is that, unlike TreeTagger, the tagset of Frog has a specific tag `NTYPE(eigen)`, which classifies nouns written with a capital letter (such as names of markets, people), months and days of the week as proper names. Frog does, however, classify all of the tokens which should be in that category in our corpus as `SPEC(deeleigen)`. This is the tag given to proper nouns which consist of multiple parts, such as the street name 'Jozef Paelinckstraat' (cf. the transcription protocol of the CGN).[12] Not all of the instances of the tag `SPEC(deeleigen)` were incorrectly assigned though.

3.4 Discussion

After an error analysis of Frog and TreeTagger, we see that both classifiers handle dialect lexemes fairly well, as they do not only rely on the lexemes on which they were trained, but also on the context. Most of the tokens that have been labeled wrongly are interjections. This problem can be solved by adapting the tagger to automatically tag a predefined list of interjections, but the tags could also be adapted afterwards by automatically replacing the wrong labels of interjections by feeding a list of tokens into a Python script. In both cases, the input for the tagger or the postprocessing script consists of a list of standardized interjections. Such a list is already provided to the transcribers via the transcription protocol, so whenever the protocol has been applied correctly, a semi-standardized form of these regional interjections was used in the transcription phase.

Both taggers also have problems with proper names, but their approaches differ. TreeTagger does not provide a specific tag for proper names at all. Frog

11 A further use of *en*, viz. as the masculine form of the subject pronoun in the third-person singular in the westernmost Flemish dialects, does not pose a problem for the tagger, as pronouns appear in their Standard Dutch form (*hij*) in the second transcription layer, which forms the input to the tagger.

12 http://nederbooms.ccl.kuleuven.be/documentation/manual-EN-POS-CGN.pdf

to the contrary does recognize and label all proper names, but only as part of a proper noun which consists of multiple parts, even if the proper name consists of only one word form. If we were to opt to work with Frog, most of these cases could, however, also be corrected automatically with a script which replaces the wrongly assigned tag `SPEC(eigen)` with the tag `NTYPE(eigen)` whenever no other `NTYPE(eigen)` tag precedes or follows this tag. The same approach can be taken to assign the correct tag to the wrongly tagged *neen* in Frog (labelled `N` instead of `ITJ`), as this error, too, occurs consistently. A last type of cases is the one in which we used the token ggg to indicate fragments in which the speaker coughs or laughs. These tokens and tags can easily be processed with the same script.

If we implement the postprocessing steps described above to our data, we can already optimize the results of Frog significantly. In Maldegem for instance, where Frog obtained its highest accuracy, the accuracy increases up to 98.3% (97.3% without postprocessing). With an approach of automatic POS tagging combined with a postprocessing script we get an accuracy of 96% (89.2% without postprocessing) in Sint-Joris-Weert, where Frog obtained its lowest accuracy.

The combination of the fact that the errors in Frog are easy to solve automatically, the more extended tagset containing necessary labels for our data and the fact that it reaches the highest average accuracy for POS tagging the transcription layer of the southern Dutch pilot corpus, leads us to a decision in favor of tagging with Frog, or a similar tagger with a more extensive tagset than the NlWaC POS tagset of TreeTagger. The high accuracy of such a POS tagger has the great advantage that manual corrections of wrongly-assigned tags in other contexts than the ones described above are minimal and can be made relatively quickly.

4 Parsing

In the last decades, parsed corpora (also named treebanks) have come to play a major role in historical linguistics. In such corpora, sentences are enriched with syntactic information and information about constituents, clauses, grammatical functions and relations. Treebanks are very powerful tools that make it possible to search for – even rare – syntactic patterns of any complexity through large amounts of data. Furthermore, the results obtained through corpus queries are easily reproducible and reusable for further research. Since the early nineties, many such treebanks have been developed, such as the Penn Parsed Corpora of Historical English (Marcus, Santorini & Marcinkiewicz 1993, Taylor, Marcus & Santorini 2003, Taylor 2007), the PROIEL corpus Haug & Johndal (2008), the Icelandic Parsed Historical Corpus (IcePaHC) (Wallen-

berg, Ingason, Sigurðsson & Rögnvaldsson 2011), the ISWOC corpus Bech & Eide (2014), the Tycho Brahe Corpus of Historical Portuguese (Galves 2018), the Corpus MCVF for historical French (Martineau, Hirschbühler, Kroch & Morin 2010) and the Corpus of Historical Low German (CHLG, Booth, Breitbarth, Ecay & Farasyn (2020)). The PROIEL and ISWOC corpora are dependency-based, the others are constituency-based. In principle, both types of parsing schemes are isomorphic (Taylor 2020, Haug 2015). Ideally, the parsing scheme itself is not the same as a syntactic analysis; it should aid the retrieval of structures. Therefore, it should be compatible with different research questions and frameworks in the interpretation of the data.[13] This consideration has also guided our search for a suitable parsing scheme and NLP tools for the GCND. In Section 4.1 we evaluate the potential of different existing parsing schemes and parsers for our data. Based on the advantages and disadvantages described in the literature, we motivate our decision for the best candidate to perform some initial experiments on our data in Section 4.2.

4.1 Candidates for parsing dialects of Dutch

For the GCND project, we compared different existing parsing schemes and (semi-)automatic parsers, based on a number of practical and theoretical considerations, such as the compatibility with existing (historical) corpora (such as those enumerated above) and the existence of readily available NLP tools to facilitate this labour-intensive step, particularly considering the large amount of data in need of processing (see Section 1.2). Besides the compatibility with other corpora, an important consideration is whether or not there are existing automatic NLP tools that can – again following the CLARIN philosophy of sustainability – be reused. For parsing as well as for POS tagging (see Section 3), the adaptation of existing software (possibly with manual postprocessing of the labeled data) is less time-consuming than creating and training new tools. That is, for automated parsers, we need to consider (i) the data on which the parser was originally trained, especially if the training data were close enough to the target data, (ii) how up-to-date the parser is, and whether it is possible to retrain and/or adapt it, and (iii) the amount of manual correction needed. In any case, any type of parser will require some manual correction, but the question is which parser minimizes the workload at this stage.[14]

13 For instance, the historical Penn parsed corpora do not have a separate VP-layer; all constituents of a clause are immediate daughters of the clause node. This makes it possible to find them using only precedence and dominance relations.

14 Within the GCND project, automatically assigned annotations will be checked for their accuracy and corrected by student assistants who understand (and preferably speak) the dialect

The first parsing scheme we took into consideration, given its frequent use in many of the historical parsed corpora enumerated above, is the Penn Treebank system. In most corpora, the first phase in the construction of a fully-parsed corpus is the grouping of tokens into constituents, which is referred to as shallow parsing or chunking. This phase can be automatized. In the Penn Treebank Project, in which constituents are represented with brackets, this initial bracketing was done with a deterministic parser called Fidditch (Hindle 1983), while in other Penn Treebank based corpora, different strategies are used. In the CHLG for instance, which uses largely the same parsing scheme (with minor adaptations), chunking is done with a custom-developed, rule-based Python script which starts from the already assigned POS tags and their position relative to each other. The annotators need to correct the output of this initial chunking phase manually. To this purpose, tools like Annotald (Beck, Ecay & Ingason 2015) can be used. Annotald is a mouse-based graphical user interface that allows the annotator to combine the syntactic chunks into larger clauses. An example of the Penn bracketing system in Annotald, applied to a GCND transcription using TreeTagger POS tags, is shown in Figure 4.

A big advantage of the Penn parsing scheme is that it is compatible with any type of POS tagset, adding syntactic structure on top of the POS annotation.[15] A disadvantage of this type of deep parsing is that it requires a lot of manual correction, making it too time-consuming for an extensive corpus such as the GCND. We therefore explored two other (semi-)automatic tools.

The first (semi-)automatic tool we considered is the constraint-satisfaction inference-based dependency parser (CSI-DP, Canisius, Bogers, Van den Bosch, Geertzen & Sang (2006), Canisius (2009)), which is one of the NLP tools of Frog (see Section 3.2). CSI-DP is trained on the one hand on the syntactically annotated LASSY small corpus (van Noord, Bouma, Eynde, de Kok, van der Linde, Schuurman, Sang & Vandeghinste 2013), which is parsed with the Alpino parser (see below) and manually verified. On the other hand, it is trained on several million tokens of text automatically parsed by the Alpino parser as well. CSI-DP is a dependency parser which bases its parses on the allocated POS tags using the TiMBL classifier. The tags are the same as the ones which are used by the Corpus of Gesproken Nederlands. An advantage of this parser is that if the POS tags from Frog – the best tagger for our data – can be used, the tagset does not need to be converted before parsing. Furthermore, the system is very user-friendly. Table 3 illustrates an example of the

of the transcribed recording.

15 For instance, the CHLG, which uses a customized POS tagset for Middle Low German, HiNTS (Barteld, Ihden, Dreessen & Schröder 2018), uses the same syntactic parsing scheme as the Penn Corpora of Historical English, with some minor adaptations (Booth et al. 2020).

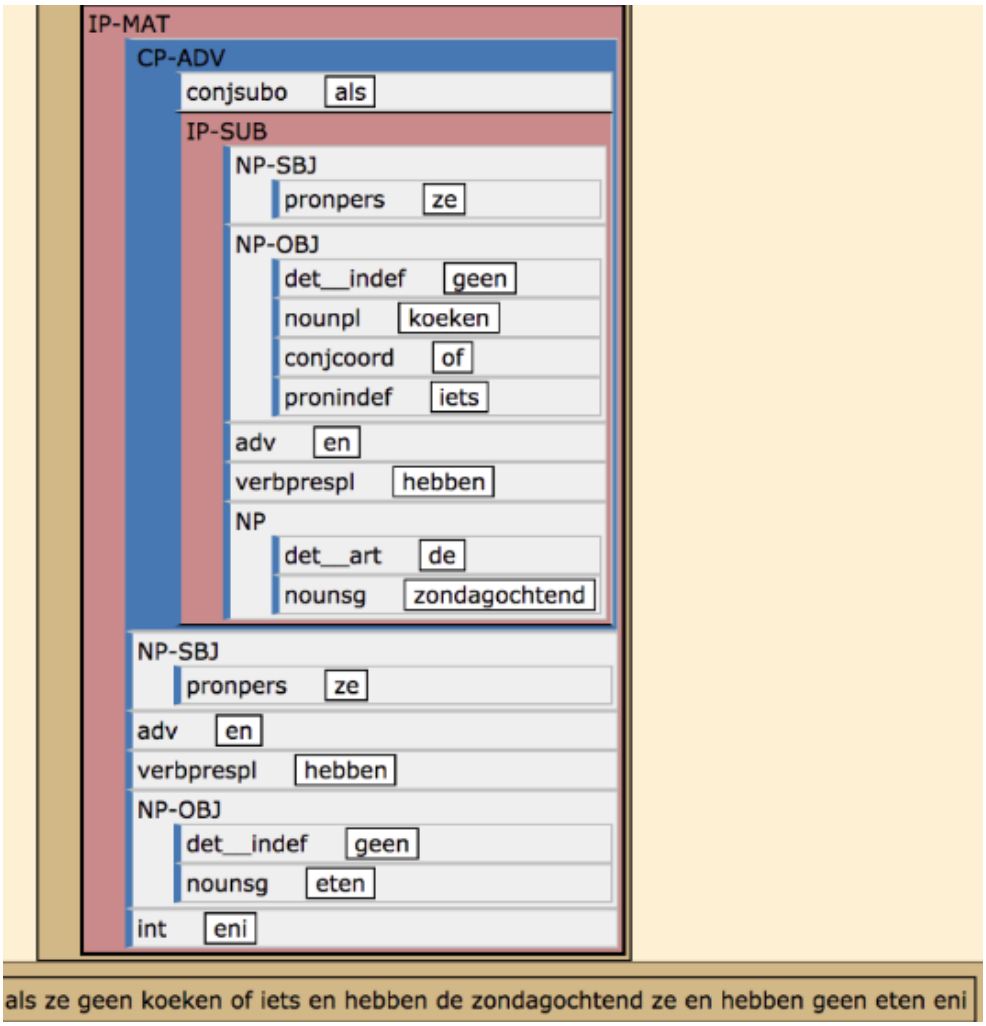

Figure 4 Editing Penn-style bracketing of GCND transcriptions in Annotald.

output of Frog's CSI-DP applied to GCND transcriptions, based on an excerpt from the transcription of Ypres (N72). A disadvantage, however, is that it is reported to perform considerably worse than Alpino, the next parser we will discuss (Van den Bosch, Busser, Canisius & Daelemans 2007). This means that using Frog's default CSI-DP would require a large amount of manual

	Token	Chunk	Token number in dependency graph	dependency relation
1	je	B-NP	2	su
2	zat	B-VP	0	ROOT
3	nog	B-ADVP	2	mod
4	in	B-PP	2	mod
5	het	B-NP	7	det
6	groen	I-NP	7	mod
7	kooltje	I-NP	4	obj1
8	.	O	7	punct

Table 3 Output of Frog's CSI-DP applied to GCND transcriptions.

correction as well.

The last parser we have considered for parsing our data is Alpino, which is a computational analyser of Dutch developed by the Computational Linguistics research group of the Center for Language and Cognition at the Rijksuniversiteit Groningen (van Noord et al. 2006). It was developed in the context of the PIONIER Project 'Algorithms for Linguistic Processing'.[16] Alpino incorporates a head-driven phrase structure grammar, which has been augmented to build dependency structures, according to the guidelines of the CGN. The training data of Alpino consist of a set of syntactic annotations, i.e. the Alpino Dependency Treebank (Van der Beek, Bouma, Malouf & van Noord 2002). This treebank contains manually corrected parses of about 7100 sentences ($\pm$145.000 words) from the newspaper part of the Eindhoven corpus (Uit den Boogaart 1975, van Grootheest 1992). The training material thus contains for the most part written Standard Dutch from the Netherlands. Although this might sound like a bad fit for the GCND data, its parsing accuracy is in general markedly higher than the accuracy in predicting syntactic labels of the default Frog parser (Van den Bosch et al. 2007). Its disadvantage compared to CSI-DP is that it has a higher memory usage and that it is slower. An advantage, however, is that Alpino requires much less manual verification than a parser in which shallow parses need to be checked and/or annotated further manually. Alpino can be used as an alternate option for CSI-DP within Frog, although it is not integrated by default and needs to be installed locally, after which the parser output can be integrated using an ex-

16 https://www.let.rug.nl/vannoord/alp/Alpino/

tra feature.[17] As we plan to retrain the parser on our dialect data in the next phase of the process, this will not be an option, as we will need to run an adapted version of the parser in a different environment.

4.2 Testing Alpino

Based on the discussion in Section 4.1, Alpino seems to be a good candidate for parsing our GCND pilot data. The reason for this is its high accuracy for parsing Standard Dutch and the limited manual labor involved in parsing beyond shallow parsing. The Alpino parser does not build on allocated POS tags or on phrase chunks, but on the token (word) itself.[18] The exported transcribed data from ELAN can be parsed without adaptation of the existing tools, although some postprocessing of the output files is needed. The input data for Alpino needs to be tokenized, with the sentences divided into one sentence (fragment) per line and each token (including punctuation) separated by a space. Furthermore, an identifier needs to be allocated to each line. For the GCND project, we made these changes using a customized Python script.

In order to identify which syntactic structures are recognized and which are not, and to see how much adaptation of the parser would be necessary, we carried out initial tests with the unadjusted parser which was only trained on the Alpino Dependency Treebank. Again, we have used transcriptions of recordings which were made in all the large dialect groups, including the transition areas between the dialects. The ten localities on which we have performed tests with the parser are the same as the ones we tested the POS taggers on (cf. Figure 3). When evaluating the output of these initial parses in our pilot corpus, a few shortcomings for our spoken material can be observed immediately. This is because the standardization we have applied to

17 https://frognlp.readthedocs.io/en/latest/advanced.html#parser

18 During the process of parsing, Alpino can also produce CGN POS tags as part of its output, based on the parse trees. That means that the accuracy of the POS tags largely depends on the accuracy of the parser itself. As a result, there are a number of problems which did not occur using the other POS taggers. For example, where the parser thinks there are hesitations or reformulations, the tags in the parse trees are placed outside the main clause and those elements are skipped by the POS tagger, while with another POS tagger, they would receive a POS tag. Because of this, we have decided not to calculate an accuracy for the Alpino POS tags in this phase of the project. An advantage of choosing Alpino for POS tagging as well as for parsing would be that the POS and parsing data would not need to be converted to other formats or integrated in other systems anymore to link them to each other. Furthermore, Alpino also uses the extensive CGN tagset which was used by the Frog POS tagger as well. Although we can only evaluate Alpino's POS tags in the next phase of the project once the parser has been optimized, we expect similar advantages and disadvantages as the ones described for the POS tags assigned by Frog.

the second layer of the transcription does not apply to word order and dialect syntax in general. Since the intention of the GCND is to facilitate syntactic research, it is of course of the utmost importance that the word order from the dialects is preserved. For this reason, it is inevitable that certain structures are uninterpretable for the parser that so far has only been trained on Standard Dutch. We see for instance that the parser will have to be adjusted primarily with respect to (i) (morpho)syntactic structures which do not occur in Standard Dutch (see Section 4.3) and (ii) very long and/or heterogeneous units of information. The latter are typical for spoken language and include, for instance, sentences in which the speaker has to search for words (8 a), sentences in which there are hesitations or in which the speaker reformulates something (8 b) or sentences in which an interjection, a short sentence or a phrase interrupting a running sentence is inserted (8 c).

(8) a. *en we hebben geraakt euh* ***n...*** *tussen Soissons en*
and we have reached ITJ n... between Soissons and
Villers-Cotterêts.
Villers-Cotterêts
'And we got there between n... Soissons and Villers-Cotterêts.'
(S020a Buysscheure)

b. ***als we...*** *eu over tijd het was hier volk eu die je*
if we ITJ over time it was here people ITJ REL you
kon je werk doen doen en eu je moest je werk
could your work do do and ITJ you must your work
schikken om het heel jaar lang je volk te kunnen
arrange to the whole year long your people to can
eu gebruiken
ITJ use
'When we... Over time, there were people here who you could hire to do your work and you had to arrange your work to be able to use your people all year round.' (N108 Hondegem)

c. *het is daar zelfs eu nog eu al de andere kant van dat*
it is there even ITJ still ITJ at the other side of that
water een manier van eu ***ah kijk ik kan dat niet juist***
water a sort of ITJ o look I can that not right
zeggen in het Vlaams ik ga het zeggen gow of in het
say in the Flemish I go it say ITJ or in the
Frans *un tour* ***het is te zeggen*** *een kot een prison.*
French a prison it is to say a kennel a prison
'At the other side of the water there is even some sort of... O look, I

> can't say it in Flemish - I am going to say - or in French... a prison - it is to say - a kennel... a prison'. (N108 Hondegem)

Van Noord et al. (2006) describe how the accuracy of the Alpino parser greatly reduces when the sentence under consideration includes more than twenty tokens. This will form a serious challenge for the rest of the parsing of our corpus, which we plan for the next phase of the project. The anomalous syntactic structures will require running the parser on a sample of example sentences containing these specific structures (van Noord et al. 2006). Very low-frequent structures, however, might require manual checking and postprocessing.

In order to discover the sentences the parser has problems with, van Noord et al. (2006) describe how an error-mining technique can be used to discover systematic problems which cause the parser to fail, resulting in a parsability table containing an overview of all those problems. The development of such a parsability table for our data would be promising, as it could shed light on common problems or (morpho)syntactic structures deviating from Standard Dutch, which might not have been (extensively) described yet. However, in order to do this, a much larger dataset than the one we have at the moment will be needed, which will be realized in a next phase of the project. In the next Section, we provide a few first case studies based on some syntactic features which we came across in the first parsing results when checking the output manually.

4.3 Case studies

In this Section, we exemplify the type of problems in the first parsing results with two case studies.

4.3.1 V>2

One of the syntactic structures which are not correctly parsed by Alpino are V>2 structures. They rarely occur in Standard Dutch, but are typical for the most western dialects in the corpus. Dutch and all other Germanic standard languages, with the exception of modern English, have V2 word order in a declarative main clauses (example 9 a), i.e. the finite verb linearly takes the second place in the sentence. After a preceding constituent, e.g. after an adverb, inversion occurs (example 9 b).

(9) a. *Het spatte in alle richtingen.*
it splashed in all directions
'It splattered in all directions.'

b. *Als je erop sloeg met een pikhouweel, spatte het.*
if you on-it hit with a pickaxe splattered it
'It you hit on it with a pickaxe, it splattered'

The situation is different in some SDDs. In West Flemish in particular, sentences with initial adverbial constituents often occur without inversion of verb and subject (Vanacker 1977, Lybaert, De Clerck, Saelens & De Cuypere 2019, Haegeman & Greco 2018). In those sentences, which seem to violate the V2 condition, the subject immediately follows the preceding phrase (example 10). In French Flemish, the left periphery can be even more complex, as can be seen in (11), where *'t maakt* does not function as a main clause followed by a complement clause, but rather as a discourse marker introducing a new topic (Farasyn 2021).[19]

(10) *a je derop sloeg met een pioche het spetterde*
if you on-it hit with a pickaxe it splattered
'It you hit on it with a pickaxe, it splattered' (H116p Torhout)

(11) *'t maakt de boerinnen als ze anders kunnen leven*
it makes the peasant.women if they otherwise could live
ze laten 't ook vallen eneeë
they let it also fall ITJ
'If the peasant women can live otherwise, they also stop doing it, isn't it?' (S010p Loberghe)

At the moment, in examples similar to (11), one of the main problems is that Alpino returns as best parse a structure in which *het*/*'t* is the subject of the clause, *maakt* is the syntactic head and *de boerinnen* is the direct object. In examples similar to (10), however, the parser often correctly analyses the subject and the other elements in the main clause, and the sentence-initial adverbial constituent is indeed correctly analysed as a separate constituent, all belonging to the same root node of the dependency structure. In many other similar cases, however, the structure is analysed incorrectly. As structures such as (10) and (11) are relatively common, the latter especially in

19 Cf. similar discourse markers developing out of former matrix clauses e.g. *ich mein* 'I mean' in German (Günthner & Imo 2003).

French Flanders, the solution will lie in retraining the parser on a set of similar sentences, which can – because of its high frequency – easily be compiled manually from the data which have been transcribed so far.[20]

4.3.2 Non-negative 'en'

The particle *en* was already mentioned in Section 3.3.2, as it is usually classified as the conjunction *en* 'and'.[21] In many contexts, however, *en* is actually the original marker of negation, which was preserved in several Flemish varieties as a remnant of Jespersen's cycle (Koelmans 1967, Neuckermans 2008, Haegeman et al. 1995). According to Neuckermans (2008), *en* appears more often in negative subclauses than in negative main clauses, normally directly in front of the finite verb. Furthermore, she gives examples of restrictive and expletive uses, but also of rare cases of non-negative uses, which seem to be limited mainly to Brabant (and a place in East Flanders), where *en* can also precede a nonfinite verb.

In the GNCD currently under construction, there are already indications of a much more far-reaching use of non-negative *en*. In the pilot corpus, it appears for instance in a (non-negative) subclause of a negative main clause, in a (non-negative) subclause of a main clause with a restrictive particle (*maar* 'but'), and as the only marker of negation in a main clause with a interrogative complement clause, a construction described mainly for Middle Dutch (Postma (2002), example 12 a). Moreover, it is attested in non-negative sentences, and preceding non-finite verbs (example 12 b), which might indicate that there is an ongoing change in the use of (non-)negative *en*.

(12) a. *ik* ***en*** *weet of dat nu nog veel meer gedaan werd*
I EN know if that now still much more done gets
'I don't know if that is still often done nowadays' (O265 Ronse)

b. *met zijn beste kleren aan... je had die een keer*
with his best clothes on... you had that.one a time
moeten en zien
must EN see

20 Note that this does not mean that a syntactic analysis of such structures is imposed on the parser. Rather, this is intended to improve the performance of the parser, to make the retrieval of such structures, which deviate from Standard Dutch, possible at all. As alluded to above, the purpose of syntactic parsing is not syntactic analysis; the parsing scheme should ideally be theory-neutral. The theoretical interpretation of the data retrieved from a parsed corpus is then the task of the linguist.

21 Both particles, while now homophonous, have different historical sources. Negative *en* derives from the Germanic negator *ni*; the coordinating conjunction *en* derives from *inde*.

'with his best clothes on... you should have seen that one!'
(N42 Pittem)

At the moment, the parser (and the tagger) consistently tag(s) *en* as the coordinating conjunction. Given the different syntactic distribution of the conjunction and the preverbal particle, the strategy will be to retrain the parser on the basis of manually corrected examples (with *en* in the relevant cases POS-tagged as an adverb) with the aim to improve the final parsing results.

5 Conclusion

This paper reported on the construction of the GCND, a tagged and parsed corpus of the spoken SDDs. We presented the results of initial tests reusing NLP tools in the light of the CLARIN philosophy. The NLP tools we have tested (i.e. Frog, TreeTagger and Alpino) have been optimized or developed for POS tagging and/or parsing Standard Dutch and were applied to a set of transcriptions representative for the southern Dutch language area under consideration. The findings of the experiments described in this paper will inform the further construction of a larger tagged and parsed corpus of the SDDs, the GCND, which will contain over 750 recordings or about 570 hours of spoken dialect data facilitating diachronic syntactic research of Dutch.[22]

The tested POS taggers both had an average accuracy of over 90%, which is relatively high for historical corpora of non-standardized languages. This is mainly due to the fact that we chose to transcribe in two layers and to apply the POS tagger on the layer which is closer to Standard Dutch. The problems with the POS tagger mainly concerned wrongly classified interjections, which frequently occur in the dialects. In the future, we plan to overcome this by automatically correcting the tags by means of a postprocessing script based on a predetermined list of interjections. We described how similar measures can be taken to postprocess wrongly allocated tags automatically, for instance for named entities, negation particles etc. The tests furthermore showed that the dialect data in the corpus require a POS tagger which uses a rather extensive tagset, such as the Frog POS tagger which uses the CGN tagset.

For the evaluation of existing parsers, we made our choice mainly based on the suitability of the tagset to our data, the training data, the possibility to adapt the existing parser, and the amount of manual verification and/or postprocessing needed, which led us to choose the Alpino parser for the future enrichment of the corpus. There are two main types of problems that still have to be solved by modifying this parser. The first is the occurrence

22 https://www.gcnd.ugent.be/en/home/

of (morpho)syntactic structures which do not occur in Standard Dutch. This can be addressed by retraining the parser on a set of sentences containing such structures. The second are the sometimes very extensive units of information. These will require manual correction, as the parser struggles with longer units of information when parsing Standard Dutch as well. However, compared to the efforts needed to develop a new parser geared towards spoken dialects from scratch, the adaptation of existing parsers to historical and spoken data is in our view certainly a much more efficient approach. In line with the CLARIN philosophy, which encourages the sharing, use and sustainability of tools for research in the humanities and social sciences, we hence opt for an adaptation of tools that were initially geared at standard language annotation. It is in our view also essential that researchers report elaborately on (i) the paths explored to annotate their data – even paths that turned out to be less successful – and (ii) the amount and kind of manual postprocessing needed. Such reporting is after all very valuable for researchers involved in future annotation projects. With this paper, we hope to have offered inspiration for researchers exploring annotation methods for non-standard spoken language, especially non-standard Dutch.

References

Barbiers, Sjef & Hans Bennis. 2007. The syntactic atlas of the Dutch dialects. *Nordlyd* 34(1). 53–72. http://septentrio.uit.no/index.php/nordlyd/article/view/89. DOI: https://doi.org/10.7557/12.89.

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Johan van der Auwera & Margreet van der Ham. 2008a. *Syntactische Atlas van de Nederlandse Dialecten. Deel II*. Amsterdam University Press.

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos & Margreet van der Ham. 2005. *Syntactische Atlas van de Nederlandse Dialecten. Deel I*. Amsterdam University Press.

Barbiers, Sjef, Olaf Koeneman, Marika Lekakou & Margreet van der Ham (eds.). 2008b. *Microvariation in Syntactic Doubling*. Brill.

Barteld, Fabian, Sarah Ihden, Katharina Dreessen & Ingrid Schröder. 2018. HiNTS: A Tagset for Middle Low German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 3940–3945.

Bech, Kristin & Kristine Eide. 2014. The ISWOC corpus. http://iswoc.github.io.

Beck, Jana, Aaron Ecay & Anton Karl Ingason. 2015. Annotald. Version 1.3. 7.

Van der Beek, Leonoor, Gosse Bouma, Rob Malouf & Gertjan van Noord.

2002. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands 2001*, 8–22. Brill Rodopi.

Blancquaert, Edgard. 1948. Na meer dan 25 jaar dialect-onderzoek op het terrein. *Koninklijke Vlaamse Academie voor Taal- en Letterkunde* III. 5–62.

Uit den Boogaart, Pieter C. 1975. Woordfrequenties in geschreven en gesproken Nederlands. *Utrecht: Oosthoek, Scheltema en Holkema* .

Booth, Hannah, Anne Breitbarth, Aaron Ecay & Melissa Farasyn. 2020. A Penn-style Treebank of Middle Low German. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)* 766–775.

Brandner, Ellen. 2015. Syntax des Alemannischen (SynAlm). Tiefenbohrungen in einer Dialektlandschaft. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (eds.), *Regionale Variation des Deutschen. Projekte und Perspektiven*, 289–322. De Gruyter. https://www.ling.uni-stuttgart.de/institut/ilg/forschung/projekte/synalm/.

Breitbarth, Anne & Liliane Haegeman. 2014. The distribution of preverbal *en* in (West) Flemish: syntactic and interpretive properties. *Lingua* 147. 69–86.

Canisius, Sander, Toine Bogers, Antal Van den Bosch, Jeroen Geertzen & Erik Tjong Kim Sang. 2006. Dependency parsing by inference over high-recall dependency predictions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, 176–180.

Canisius, Sander Valentijn Maria. 2009. *Structured prediction for natural language processing: A constraint satisfaction approach*: Tilburg University dissertation.

Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Kurt Van Der Sloot & Antal Van den Bosch. 2004. Timbl: Tilburg memory-based learner. *Tilburg University* .

De Vogelaer, Gunther & Magda Devos. 2008. On geographical adequacy, or: How many types of subject doubling in Dutch. In *Microvariation in syntactic doubling*, 251–276. Brill.

Farasyn, Melissa. 2021. V(>)2 in de declaratieve hoofdzin in de Frans-Vlaamse dialecten. *Handelingen Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis* 74. 81–97.

Fleischer, Jürg, Alexandra N. Lenz & Helmut Weiß. 2017. SyHD-Atlas. Konzipiert von Ludwig M. Breuer. Unter Mitarbeit von Katrin Kuhmichel, Stephanie Leser-Cronau, Johanna Schwalm und Thomas Strobel. http://www.syhd.info.

Galves, Charlotte. 2018. The Tycho Brahe Corpus of Historical Portuguese:

Methodology and results. *Linguistic Variation* 18(1). 49–73.

Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen & Arjan van Hessen. 2020a. Clearing the transcription hurdle in dialect corpus building: The corpus of Southern Dutch dialects as case-study. *Frontiers in Artificial Intelligence* 3. 1–17. DOI: 10.3389/frai.2020.00010.

Ghyselen, Anne-Sophie, Jacques Van Keymeulen, Melissa Farasyn, Lien Hellebaut & Anne Breitbarth. 2020b. Het transcriptieprotocol van het Gesproken Corpus van de Nederlandse Dialecten (GCND). *Bulleting de la Commission Royale de Toponymie & Dialectologie / Handelingen van de Koninklijke Commissie voor Toponymie & Dialectologie* 92. 83–115. DOI: 10.2143/TD.92.0.3288999.

Ghyselen, Anne-Sophie & Jacques Van Keymeulen. 2014. Dialectcompetentie en functionaliteit van het dialect in Vlaanderen anno 2013. *Tijdschrift voor Nederlandse taal-en letterkunde* 130(2). 117–139.

van Grootheest, Dave. 1992. Handleiding bij het eindhoven corpus (vu-versie). *Rapport technique, Vrije Universiteit Amsterdam* .

Günthner, Susanne & Wolfgang Imo. 2003. Die Reanalyse von Matrixsätzen als Diskursmarker: *ich mein*-Konstruktionen im gesprochenen Deutsch. In Magdolna Orosz & Andreas Herzog (eds.), *Jahrbuch der ungarischen Germanistik*, 181–216. Gesellschaft ungarischer Germanisten / DAAD.

Haegeman, Liliane. 1992. *Theory and description in generative syntax: A case study in West Flemish*. Cambridge University Press Cambridge.

Haegeman, Liliane & Ciro Greco. 2018. West Flemish V3 and the interaction of syntax and discourse. *The Journal of Comparative Germanic Linguistics* 21(1). 1–56.

Haegeman, Liliane & Virginia Hill. 2013. The syntacticization of discourse. *Syntax and its limits* 48. 370–390.

Haegeman, Liliane & Marjo Van Koppen. 2012. Complementizer agreement and the relation between C^0 and T^0. *Linguistic Inquiry* 43(3). 441–454.

Haegeman, Liliane et al. 1995. *The syntax of negation*. Cambridge University Press.

Haug, Dag. 2015. Treebanks in historical linguistic research. In Carlotta Viti (ed.), *Perspectives on historical syntax*, Benjamins.

Haug, Dag T. T. & Marius L. Johndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.

Herrgen, Joachim. 2010. The digital Wenker Atlas (www.diwa.info): An online research tool for modern dialectology. *Dialectologia* Special Issue I.

89–95.

Hindle, Donald. 1983. User manual for Fidditch, a deterministic parser. *Naval Research Laboratory Technical Memorandum* 7590. 142.

de Jong, Franciska, Bente Maegaard, Koenraad de Smedt, Darja Fišer & Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 3259–3264.

Kehrein, Roland, Alfred Lameli & Stefan Rabanus (eds.). 2015. *Regionale Variation des Deutschen: Projekte und Perspektiven / Regional Variation in German: Projects and Perspectives*. De Gruyter. DOI: https://doi.org/10.1515/9783110363449.

Koelmans, Leendert. 1967. Over de verbreiding van het ontkennende *en*. *De nieuwe taalgids* 60(1). 12–18.

Koleva, Mariya, Melissa Farasyn, Bart Desmet, Anne Breitbarth & Véronique Hoste. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics* 22(1). 107–140.

Lindstad, Arne Martinus, Anders Nøklestad, Janne Bondi Johannessen & Oystein Alexander Vangsnes. 2009. The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages. In Kristiina Jokinen & Eckhard Bick (eds.), *NODALIDA 2009 Conference Proceedings*, 283–286. http://www.tekstlab.uio.no/nota/scandiasyn/index.html.

Lybaert, Chloé, Bernard De Clerck, Jorien Saelens & Ludovic De Cuypere. 2019. A corpus-based analysis of V2 variation in West Flemish and French Flemish dialects. *Journal of Germanic Linguistics* 31(1). 43–100.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2).

Martineau, France, Paul Hirschbühler, Anthony Kroch & Yves Charles Morin. 2010. Corpus MCVF annoté syntaxiquement. Ottawa: University of Ottawa. http://www.arts.uottawa.ca/voies/corpus_pg_en.html.

Marynissen, Ann & Guy Janssens. 2013. A regional history of Dutch. In Frans Hinskens & Johan Taeldeman (eds.), *Dutch*, vol. 3 Language and Space, 81–100. Berlin: De Gruyter Mouton.

Meelen, Marieke. 2020. Annotating Middle Welsh: POS tagging and chunk-parsing a partial corpus of native prose. *Proceedings of the Chronologicon Hibernicum workshop on Morphological and Syntactic Variation* 27–47. https://doi.org/10.1515/9783110680744-003.

Neuckermans, Annemie. 2008. *Negatie in de Vlaamse dialecten volgens de gegevens van de Syntactische Atlas van de Nederlandse Dialecten (SAND)*:

Ghent University dissertation.

van Noord, Gertjan, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang & Vincent Vandeghinste. 2013. Large scale syntactic annotation of written dutch: Lassy. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for dutch. theory and applications of natural language processing*, 147–164. Springer. https://doi.org/10.1007/978-3-642-30910-6_9.

Overdiep, Gerrit Siebe. 1933. Dialectstudie en syntaxis. *Onze Taaltuin* 2. 18–23.

Poletto, Cecilia & Paola Benincà. 2007. The ASIS enterprise: A view on the construction of a syntactic atlas for the Northern Italian dialects. *Nordlyd* 34(1). 35–52. http://septentrio.uit.no/index.php/nordlyd/article/view/88. DOI: https://doi.org/10.7557/12.88.

Postma, Gertjan. 2002. Negative polarity and modality in Middle Dutch *ghe*-particle constructions. In Sjef Barbiers, Frits Beukema & Wim van der Wurff (eds.), *Modality and its interaction with the verbal system*, 205–244. Benjamins.

Ryckeboer, Hugo. 2013. A West Flemish dialect as a minority language in the north of France. In Frans Hinskens & Johan Taeldeman (eds.), *Dutch*, vol. 3 Language and Space, 782–800. Berlin: De Gruyter Mouton.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In D.B. Jones & H. Somers (eds.), *New methods in language processing*, 154–164.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds.), *Natural language processing using very large corpora*, 13–25. Springer.

Sloot, Ko Van Der, Iris Hendrickx, Maarten van Gompel, Antal Van den Bosch & Walter Daelemans. 2018. Frog, A Natural Language Processing Suite for Dutch. Reference Guide. Tech. Rep. Language and Speech Technology Technical Report Series 18-02 Radboud University. https://frognlp.readthedocs.io/en/latest/.

Taeldeman, Johan. 2001. De regenboog van de Vlaamse dialecten. In Johan De Caluwe, Magdalena Devos & Johan Taeldeman (eds.), *Het taallandschap in Vlaanderen*, 49–58. Academia Press.

Taylor, Ann. 2007. The York—Toronto—Helsinki parsed corpus of Old English prose. In Karen P. Corrigan & Adam Mearns (eds.), *Creating and digitizing language corpora*, 196–227. Springer.

Taylor, Ann. 2020. Treebanks in Historical Syntax. *Annual Review of Linguistics* 6(1). 195–212.

Taylor, Ann, Mitchell Marcus & Beatrice Santorini. 2003. The Penn treebank:

An overview. In Anne Abeillé (ed.), *Treebanks. Building and Using Parsed Corpora*, 5–22. Springer.

Van den Bosch, Antal, Bertjan Busser, Sander Canisius & Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series* 7. 191–206.

van Noord, Gertjan et al. 2006. At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13^{e} conference sur le traitement automatique des langues naturelles*, 20–42.

Vanacker, Valeer F. & Georges De Schutter. 1967. Zuidnederlandse dialekten op de band. *Taal en Tongval* 19. 35–51.

Vanacker, Valere F. 1977. Syntactische overeenkomsten tussen Frans-Vlaamse en Westvlaamse dialekten. *De Franse Nederlanden. Les Pays Bas français. Jaarboek. Ons Erfdeel* 206–216.

Vandekerckhove, Reinhild. 2009. Dialect loss and dialect vitality in Flanders. *International Journal of the Sociology of Language* 196/197. 73–97. DOI: 10.1515/IJSL.2009.017.

Van der Wal, Marijke. 1995. De moedertaal centraal. Standaardisatie-aspecten in de Nederlanden omstreeks 1650. *Nederlandse cultuur in Europese context, monografieën en studies* .

Wallenberg, Joel C, Anton Karl Ingason, Einar Freyr Sigurðsson & Eiríkur Rögnvaldsson. 2011. Icelandic parsed historical corpus (IcePaHC). *Version 0.9* http://www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC).

Zampieri, Marcos, Preslav Nakov, Shervin Malmasi, Nikola Ljubešić, Jörg Tiedemann & Ahmed Ali (eds.). 2019. *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. Ann Arbor, Michigan: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-1400.

Melissa Farasyn
Blandijnberg 2
9000 Gent
Belgium
Melissa.Farasyn@UGent.be
https://research.flw.ugent.be/en/melissa.farasyn

Anne-Sophie Ghyselen
Blandijnberg 2
9000 Gent
Belgium
AnneSophie.Ghyselen@UGent.be
https://research.flw.ugent.be/en/annesophie.ghyselen

Jacques Van Keymeulen
Blandijnberg 2
9000 Gent
Belgium
Jacques.VanKeymeulen@UGent.be
https://research.flw.ugent.be/en/jacques.vankeymeulen

Anne Breitbarth
Blandijnberg 2
9000 Gent
Belgium
Anne.Breitbarth@UGent.be
https://research.flw.ugent.be/en/anne.breitbarth