CHAPTER 22

GrETEL
A Tool for Example-Based Treebank Mining

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman and Frank Van Eynde
Centre for Computational Linguistics, KU Leuven

ABSTRACT

This chapter describes the use of GrETEL for linguistic research. GrETEL is a linguistic search tool that enables users to look up constructions in syntactically annotated corpora or *treebanks*. It provides online access to the data, allowing users to query a treebank using either an example sentence or an XPath expression in order to look for similar constructions. A major asset of GrETEL is that it enables non-technical users to consult treebanks in a user-friendly way, which is also in line with the main CLARIN goal of applying the results of speech and language technology to research in the humanities and the social sciences. Besides a description of the querying procedure in GrETEL, this chapter presents a selection of research in Dutch syntax and semantics that has been carried out using GrETEL. Furthermore, an overview is given of further developments.

22.1 Introduction

The construction of syntactically annotated corpora or *treebanks* has created exciting opportunities for the empirical investigation of syntax.[1] For Dutch, several treebanks are available, e.g. the CGN treebank (van der Wouden et al., 2002) for spoken Dutch, and LASSY (van Noord et al., 2013) and SoNaR (Oostdijk et al., 2013) for written Dutch. While treebanks have the potential to be an added value for descriptive and theoretical linguistics, the exploitation of such treebanks usually requires that the user have in-depth knowledge of the annotation guidelines and master a formal

[1] We use the term *treebank* to refer to both manually constructed syntactically annotated corpora and automatically parsed corpora.

How to cite this book chapter:
Augustinus, L, Vandeghinste, V, Schuurman, I and Van Eynde, F. 2017. GrETEL: A Tool for Example-Based Treebank Mining. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 269–280. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.22.

query language. Some users are not deterred by this, but many are, so that the potential of the treebanks will not be realised. To make the treebanks useful for the computationally less inclined we have developed GrETEL, a user-friendly search engine for treebanks (Augustinus et al., 2012; Augustinus et al., 2013). It offers the possibility to provide the system with an example sentence in order to collect relevant corpus data. Therefore, the development of GrETEL paves the way for combining treebank mining with descriptive and theoretical linguistics.

22.2 What is GrETEL?

GrETEL stands for *Greedy Extraction of Trees for Empirical Linguistics*. It is a linguistic search engine that enables users to extract information from treebanks in a user-friendly way. Instead of a formal search instruction, it takes a natural language example as input. This provides a convenient way for novice and non-technical users to use treebanks with a limited knowledge of the underlying syntax and formal query languages.

Since linguists tend to start their research from example sentences, example-based querying allows them to use those examples as a starting point for treebank search. Work related to our approach is the Linguist's Search Engine (Resnik and Elkiss, 2005), a tool that also made use of example-based querying, but is no longer available, and the TIGER Corpus Navigator (Hellmann et al., 2010), which is a Semantic Web system used to classify and retrieve sentences from the TIGER corpus on the basis of abstract linguistic concepts.

The system we present here is an online system,[2] which shares the advantages of tools like TüNDRA (Martens, 2013) and INESS-Search (Meurer, 2012): they are platform-independent and no local installation of the treebanks is needed. This is especially attractive for (very) large parsed corpora which require a lot of disk space. Another related tool is the more recently constructed PaQu application (Odijk, 2015, see chapter 23). In addition to an online search interface, PaQu also offers the possibility to upload and parse a locally installed corpus.

For a presentation of the way in which GrETEL works we first focus on the basic search mode of example-based querying (section 22.2.1) and then we turn to more advanced modes of querying (section 22.2.2).

22.2.1 Example-Based Querying

The example-based querying procedure consists of six steps.

1. Example The user provides an example sentence, containing the syntactic construction (s)he is looking for. For instance, in colloquial Dutch the complementizer *van* 'of' is sometimes used in constructions reflecting direct speech (Coppen, 2010; Hoekstra, 2010). An example is given in (1).

(1) Hij dacht van ik zal dat morgen wel doen.
he thought of I will that tomorrow rather do
'He thought: I will do that tomorrow'.

2. Parse GrETEL automatically parses the input construction using the Alpino parser (van Noord, 2006), and returns it as a syntax tree (see Figure 22.1). The user can verify the parse tree. If Alpino returns an erroneous parse, the user is advised to choose another input example.

3. Selection matrix In the selection matrix, shown in Figure 22.2, the user indicates which parts of the entered example are relevant for the construction under investigation, as well as their level of abstraction. We have indicated lemma for *van* 'of', and word class of the verbs *dacht* 'thought' and

[2] http://gretel.ccl.kuleuven.be

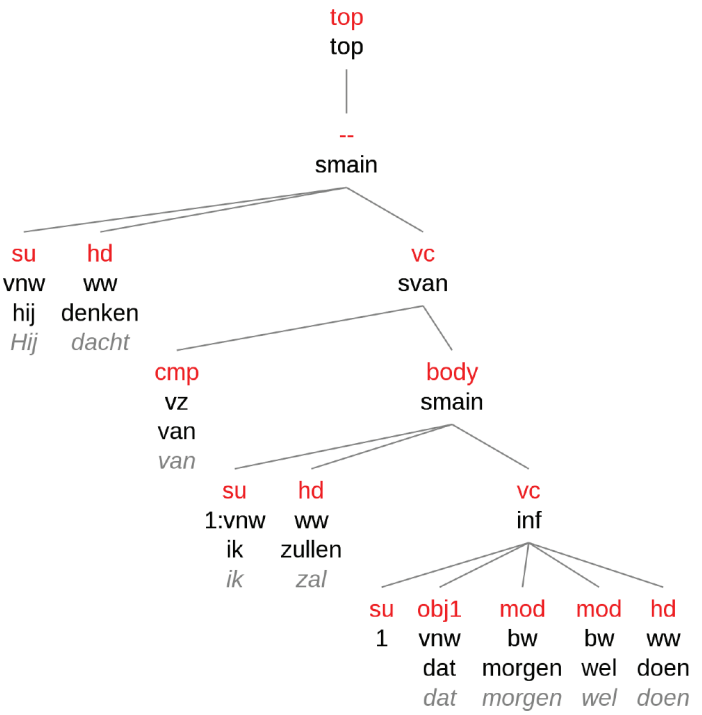

Figure 22.1: Parse tree of the input construction.

sentence	Hij	dacht	van	ik	zal	dat	morgen	wel	doen
word	○	○	○	○	○	○	○	○	○
lemma	○	○	◉	○	○	○	○	○	○
word class	○	◉	○	○	◉	○	○	○	○
optional in search	◉	○	○	◉	○	◉	◉	◉	◉

Figure 22.2: Selection matrix.

zal 'will', as we want to abstract over verb forms.[3] The other words in the example are not relevant for the construction under investigation, so those words are indicated as 'optional in search'.

The dependency relation and the word class (POS tag) of all selected items are automatically included in the search instruction. For instance, it will be taken into account that the word *van* is a preposition (tagged as vz) functioning as a complementizer (cmp).

4. Treebank selection In the next step the user can choose which treebank(s) to query. Currently one can choose between the CGN treebank for spoken Dutch, and LASSY Small and the SoNaR-500 treebank for written Dutch.[4] It is possible to query the CGN and LASSY Small treebanks as a whole, or one can select one or more treebank components, for instance to compare data from different genres. Because of its size (500 million words, ca 41 million sentences), it is

[3] The embedded verb is indicated in order to avoid constructions without an embedded sentence, such as *Hij dacht van wel* 'He thought so'.

[4] The SoNaR-500 treebank is a subset of the LASSY Large treebank.

only possible to query SoNaR per component. For this example we have chosen the part of SoNaR containing discussion lists (WR-P-E-A, 50 million words, ca 4.5 million sentences).

5. Query Based on the information provided in the selection matrix, GrETEL extracts a query tree from the parse tree (Figure 22.3). Besides the lexical information indicated in the selection matrix, the dependency relation (rel) and the phrasal category (cat) of the relevant nodes are included in the query tree, see (5). GrETEL automatically converts the query tree into an XPath expression,[5] which is used to search the treebank.

6. Results The results of the query are presented to the user as a list of sentences, with the matching part emphasised. The user can click on any of these sentences in order to visualise the results as syntax trees. For the query in Figure 22.3 GrETEL finds 175 results in the WR-P-E-A component of SoNaR. Some are presented in (2-4).

(2) Dat filmpje zegt bijna van: dit is de nieuwe norm.
That video.DIM says almost of this is the new norm
'That video almost says: this is the new norm.' (SoNaR, WR-P-E-A-0000850955.p.5.s.3)

(3) Na het voorprogramma had ik zoiets van IK BEN HIER WEG.
after the opening act had I something of I am here away
'After the opening act I was like I AM OUT OF HERE.' (SoNaR, WR-P-E-A-0000295207.p.1.s.1)

(4) ... maar ik dacht van, ik ga wachten voor da liedje.
... but I thought of I go wait for that song
'... but I thought, I'll wait for that song.' (SoNaR, WR-P-E-A-0000258967.p.1.s.1)

The results show the *greedy* nature of GrETEL: it not only returns constructions in which the parts of the construction indicated in the matrix are adjacent, but also returns examples in which those elements are discontinuous.[6] For instance, the finite verb *zegt* 'says' in (2) is not adjacent to *van* 'of'. Because of this discontinuity, looking for similar constructions in a *flat* (raw or POS-tagged) corpus would be much harder.

If we run the same query on less informal data, such as the component of SoNaR containing periodicals and magazines (WR-P-P-H), we only find 42 hits even though the corpus is larger

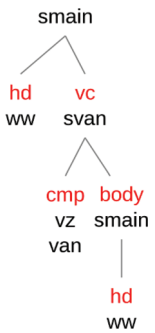

Figure 22.3: Query tree based on the input example.

[5] http://www.w3.org/TR/xpath

[6] In addition, the XPath expressions that are used (by default) in GrETEL ignore word order. This also gives rise to more general queries compared to queries used in string-based methods.

in size (ca 5.5 million sentences) than WR-P-E-A (ca 4.5 million sentences). This confirms the colloquial nature of the construction.

Example-based querying has the advantage that the user does not need to be familiar with XPath, nor with the exact syntactic structure of the XML in which the trees are represented, nor with the exact grammar implementation that is used by the parser or the annotators.

22.2.2 XPath Search

In the advanced mode of example-based search, users can inspect not only the query tree (Figure 22.3), but also the corresponding XPath expression, spelled out in (5).

(5) `//node[@cat="smain" and node[@rel="hd" and @pt="ww"] and node[@rel="vc" and @cat="svan" and node[@rel="cmp" and @pt="vz" and @lemma="van"] and node[@rel="body" and @cat="smain" and node[@rel="hd" and @pt="ww"]]]]`

Moreover, they can make modifications to this query. For instance, one can use an or-statement to construct more general queries; e.g. `node[@cat="smain" or @cat="ssub"]` looks for constructions in both main and subordinate clauses. This approach allows more flexibility in the type of patterns that are searched.

For users who are thoroughly familiar with XPath and with the details of the annotation there is also the possibility to directly formulate an XPath query describing the syntactic pattern the user is looking for. This query is then processed in the same way as the automatically generated query in the first approach.

22.3 Using GrETEL for Research and Education

GrETEL has been used for linguistic research on various topics within Dutch syntax and semantics (section 22.3.1). In addition, it has been used for teaching, and it has been presented at several conferences and guest lectures (section 22.3.2).

22.3.1 Research on Dutch Syntax and Semantics

While GrETEL has been used to investigate several linguistic topics, two strands of research received considerable attention, i.e. the investigation of verb clusters and of copular constructions.

22.3.1.1 Verb Clusters

Augustinus (2015) provides both a theoretical and a treebank-based account of Dutch verb clusters, i.e. constructions in which multiple verbs group together. She shows how such constructions can be extracted from the treebanks using GrETEL, and how the treebank observations serve as an empirical basis to verify the claims made by the theory. She conducted several case studies, such as word order variation in verb clusters, the occurrence of *Infinitivus pro Participio* (a.k.a. the IPP effect), and interruption of the cluster by nonverbal elements.

Dutch verb clusters are characterised by an unusual type of word order variation, i.e. one that does not entail a change of meaning, as shown by the examples in (6).

(6) a. ... dan denk ik ook dat die man 't heeft gedaan.
... then think I also that that man it has done
'... in that case I also think that that man has done it.' (CGN, fna000458_166)

b. ah dan hoop ik dat Ivo dat gedaan heeft.
ah then hope I that Ivo that done has
'ah in that case I hope that Ivo has done that.' (CGN, fva400092_136)

Augustinus (2015) investigates which types of word order variation occur in non-dialectal varieties of Dutch, i.e. in the CGN and LASSY Small treebanks included in GrETEL. Barbiers and Schuurman (2015) compare the word order variation in three-verb clusters encountered in those treebanks to data obtained from MIMORE, a tool for investigating morphosyntactic variation in Dutch dialects.[7]

Infinitivus pro Participio or IPP refers to constructions in which an infinitive occurs instead of a past participle, as in (7).

(7) a. Ik heb 't twee keer zien gebeuren.
I have it two times see.IPP happen
'I have seen it happen twice.' (CGN, fna000773_212)

b. * Ik heb 't twee keer gezien gebeuren.
I have it two times seen happen

IPP appears in a subset of the Germanic languages, such as Dutch, German and Afrikaans. These languages differ, however, with respect to the set of verbs that can appear as IPP verbs, and with respect to whether the phenomenon occurs obligatorily or optionally. For some verbs, the literature is not conclusive on whether they can occur in IPP constructions or not. Augustinus and Van Eynde (2012) and Augustinus (2015) describe how a treebank-supported investigation of Dutch IPP verbs using GrETEL results in a more exhaustive and empirically valid typology of Dutch IPP verbs than the lists available in the literature.

Augustinus and Van Eynde (2017) compare the set of Dutch IPP verbs to the German IPP verbs. In order to add this cross-linguistic perspective, they queried two German treebanks using the TüNDRA treebank search tool (Martens, 2013). The case study not only illustrates how the results obtained by GrETEL can be complemented by using additional resources, but also shows how the treebank data can be employed to evaluate theoretical accounts of IPP.

A third case study on IPP using GrETEL investigates the choice of the auxiliary of the perfect in Dutch IPP constructions, i.e. the choice between *hebben* 'have' and *zijn* 'be'. Canonically the choice for the auxiliary in IPP constructions is determined by the IPP verb, as in (8a). However, one also encounters constructions in which the auxiliary is determined by the main verb, as in (8b).

(8) a. en Erwin Jans die is er weer bij komen zitten want …
and Erwin Jans who is there again with come.IPP sit because …
'and Erwin Jans, he has come to join us because …' (CGN, fvl600281_1)

b. heeft er niemand dat komen zeggen tegen jullie?
has there no one that come.IPP say to you
'Did nobody come to tell you that?' (CGN, fva400386_18)

While this variation has been reported in the literature, no large-scale corpus study was available pointing out the frequency and the distribution of the phenomenon. Van Eynde et al. (2016a) investigate the choice between *hebben* 'have' and *zijn* 'be' in IPP constructions by means of GrETEL and OpenSoNaR.[8] The corpus study provides insight in the set of verbs that allow this alternation. For the verbs *moeten* 'must' and *kunnen* 'can' the distribution of the canonical and the alternative construction is investigated in more detail.

Besides word order variation and the IPP effect, Augustinus and Van Eynde (2014) and Augustinus (2015) investigate the occurrence of cluster interruption. Canonical verb clusters

[7] http://www.meertens.knaw.nl/mimore/

[8] OpenSoNaR provides string search of the flat SoNaR-500 corpus. (http://opensonar.clarin.inl.nl)

cannot be interrupted by nonverbal elements (9). There are some exceptions though, such as *cluster creeping* by separable verb particles, predicative adjectives, and stranded adpositions (10).

(9) a. ... de voorstellen die de NS vandaag heeft gedaan ...
... the proposals that the NS today has done ...
'... the proposal that the NS has made today ...' (CGN, fnk001631_2)

b. * ... de voorstellen die de NS heeft vandaag gedaan ...
... the proposals that the NS has today done ...

(10) De plicht die hem nu roept, kan hem straks de mooiste baan kosten waar een
the duty than him now calls can him later the most-beautiful job cost where a
Beier kan *van* dromen.
Bavarian can of dream
'The duty that calls him now can cost him the most beautiful job a Bavarian can dream of.'
(LASSY, WR-P-P-I-0000000033.p.21.s.4)

The treebank investigations conducted in Augustinus and Van Eynde (2014) and Augustinus (2015) show that the set of cluster creepers is larger than the literature suggests. This illustrates once more how a treebank-based investigation can provide additional insights into syntactic phenomena.

22.3.1.2 Copular Constructions

In addition to the research on verb clusters, GrETEL is used for research on copular constructions. Van Eynde et al. (2014) illustrate that the set of copular verbs discussed in traditional grammars is incomplete. Typically those grammars mention a set of 10 to 15 verbs, adding, as an afterthought, that the list is not complete. By means of GrETEL treebank data were collected in order to get a more complete and empirically motivated typology of Dutch copular constructions, which consists of at least 40 verbs. As the typology is based on linguistically motivated criteria, it can be used to complete the list of verbs by investigating a larger dataset.

Van Eynde et al. (2016b) deal with number agreement in copular constructions. Canonically, there is number agreement between the subject and the predicate nominal in Dutch copular constructions, as in (11). Mismatches are not excluded, however, as shown in (12).

(11) De volgende figuur is een eenvoudig voorbeeld ...
the next figure is a simple example ...
'The next figure is a simple example...' (LASSY, dpc-bmm-001092-nl-sen.p.5.s.1)

(12) Beide aftredende bestuurders blijven wel aandeelhouder.
both resigning directors remain POL shareholder
'Both resigning directors remain shareholder.' (LASSY, WR-P-E-I-0000 049645.p.1.s.68.2)

This research demonstrates how the data obtained from the treebanks not only provide information with respect to the frequency and the distribution of number agreement in copular constructions, but also serve as an empirical basis for a theoretical analysis. In addition, the treebank data were employed to define under which circumstances mismatches between the subject and the predicate nominal are allowed.

22.3.2 Dissemination

GrETEL is currently used in courses on descriptive linguistics, syntax and semantics, corpus analysis, and computational linguistics in order to teach students how to look up syntactic

constructions and their frequencies in a treebank without requiring them to familiarise themselves with the specifics of XPath or the specific syntax of the treebank. It teaches students about syntactic parses and treebanks by providing them easy online access to large amounts of data.

As GrETEL has a focus on user-friendliness and is freely available online, it is an example of how *Digital Humanities* applications disclose datasets and computational tools, without requiring the user to have a technical background.

GrETEL was presented to a technical audience at several conferences within the field of computational linguistics and to an audience of potential users at general linguistic conferences and doctoral schools in Flanders and the Netherlands. Those lectures typically include a tutorial demonstrating the functionality and use of GrETEL, followed by a hands-on session. In addition, some case studies are discussed, showing how the results obtained from the treebanks in GrETEL can serve as an empirical basis for research in linguistics. One of the case studies includes the combined use case of GrETEL and MIMORE (Barbiers and Schuurman, 2015). It illustrates how GrETEL and MIMORE can be used as complementary tools for studies on Dutch syntax.

22.4 Further Developments

GrETEL has been designed in such a way that it can also be used for treebanks in other languages, even if they have different annotation schemes compared to the Dutch treebanks. In the AfriBooms project (Augustinus et al., 2016a), a treebank for Afrikaans has been developed, which is also included in GrETEL (section 22.4.1). In the context of the SCATE project (Vandeghinste et al., 2016), GrETEL was adapted to query parallel treebanks (section 22.4.2). The tool is also included in Taalportaal (Landsbergen et al., 2014), an online descriptive grammar of Dutch (section 22.4.3).

22.4.1 GrETEL for Afrikaans

In comparison to Dutch, Afrikaans is a low-resource language, so until recently no treebanks for Afrikaans were available. In the AfriBooms project a (small) treebank containing ca 50K words has been developed, based on the corpus of the South African National Centre for Human Language Technologies (NCHLT). The annotations of the treebank are manually corrected, which makes it a reliable resource for linguistic research. In addition, a first parser for Afrikaans was developed. Both the treebank and the parser are included in a version of GrETEL for Afrikaans (Augustinus et al., 2016a).[9]

22.4.2 Querying Parallel Treebanks with Poly-GrETEL

In the context of the SCATE project, large-scale parallel treebanks are constructed which are used for syntax-based machine translation. Since parallel treebanks are a valuable resource for translators and linguists as well, Poly-GrETEL was developed, i.e. an extension of GrETEL for querying parallel treebanks (Augustinus et al., 2016b).[10] Currently it contains the (automatically annotated) Europarl parallel treebank for Dutch and English.

The Europarl parallel treebank We have made an update of the treebank described in Kotzé et al. (2016): we used the data from Europarl version 7 (Koehn, 2005) and extracted the Dutch and English sentence-aligned data from `www.statmt.org`. The Dutch side was parsed with the

[9] `http://gretel.ccl.kuleuven.be/afribooms`
[10] `http://gretel.ccl.kuleuven.be/poly-gretel`

Alpino parser and the English side with the Stanford parser (Klein and Manning, 2003) with added dependencies (de Marneffe et al., 2006). The phrase structure output of the Stanford parser is converted into an XML-tree,[11] analogous to the XML-output of Alpino, as shown in Figure 22.4. Besides the syntactic annotations the parallel treebank contains node alignments.[12]

Poly-GrETEL In combination with the example-based query functionality, Poly-GrETEL avoids the need for users to be familiar with the query language and the structure of the trees in the source and target language, thus facilitating the use of parallel corpora for comparative linguistics and translation studies.

The user can query the treebanks in a similar way as in the monolingual GrETEL environment, i.e. example-based or by means of an XPath query. The main difference is that the user can choose between a bilingual and a monolingual input. In the bilingual search option the user provides two input constructions: one in English and one in Dutch. Poly-GrETEL returns two parses, and the user can indicate the relevant parts of both the English and the Dutch input examples. Poly-GrETEL automatically extracts a search instruction in a similar fashion as the monolingual GrETEL, but provides the option to return only the constructions in which the English and the Dutch query trees are aligned. It is a syntactic concordancer for parallel treebanks, as it shows how a Dutch syntactic construction is translated in English (or vice versa). One could, for instance, investigate how the Dutch *van*-construction presented in section 22.2.1 is translated in English. This makes the tool interesting not only for research in (comparative) linguistics and translation studies, but also to serve as a tool for computer-aided translation for translators and language learners.

Adding the parallel English-Dutch treebank furthermore implies that GrETEL also includes English data. Since it is possible to query the English side of the parallel treebank in a monolingual way, one can use these data for a monolingual treebank investigation of syntactic phenomena in English.

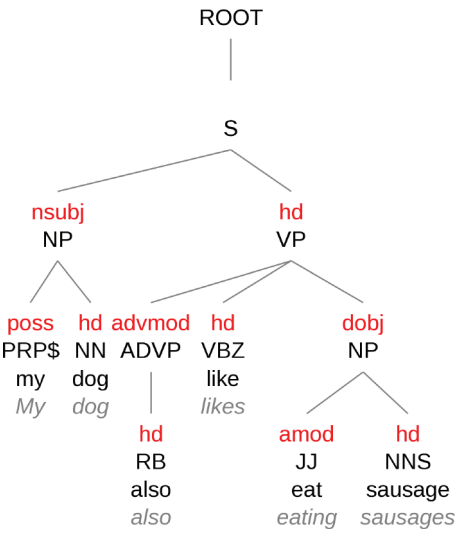

Figure 22.4: Stanford parse with added dependencies, converted into Alpino-XML.

[11] The bracketed tree and the XML tree are isomorphous.

[12] In future versions alignments resulting from several different alignment algorithms will be made available. In the current version, only alignment according to Zhechev (2009) is available.

22.4.3 Link with Taalportaal

Recently, GrETEL was linked to *Taalportaal*, a website that contains online descriptive grammars for Dutch, Frisian and Afrikaans (Landsbergen et al., 2014, see chapter 24).[13] By means of intelligent links, users can look up linguistic phenomena described in Taalportaal in a variety of online corpora, amongst others the treebanks included in GrETEL (van der Wouden et al., 2015).

The link with Taalportaal enhances the visibility of GrETEL, and encourages its use, alone or in combination with other corpus tools. Bouma et al. (2015) mention how they have used the example-based input method of GrETEL to facilitate query formulation. It turns out to be particularly useful if one does not know exactly how certain phenomena are annotated in the treebanks. In addition, the authors mention how they have used GrETEL's example-based querying functionality to become aware of differences between the treebank annotations and the analyses of the descriptive grammar included in Taalportaal (Bouma et al., 2015: 18).

22.5 Conclusion and Future Work

We have described GrETEL, a user-friendly search tool for treebanks. It originated in the context of a CLARIN-Flanders project, which aimed at the creation of tools for the exploitation of Dutch treebanks. In follow-up research, GrETEL was extended to other languages (Afrikaans and English), and other types of treebanks, i.e. parallel ones. The extensions make the tool also useful for a larger (CLARIN) audience, i.e. researchers who are not (only) working on Dutch.

Future work includes adding more languages to GrETEL, such as German and French, as for those languages we also have high-quality parsers and treebanks available.

In the framework of the Dutch CLARIAH infrastructure project and the Anncor project (University of Utrecht), there are plans to further extend the functionality of GrETEL. An upload function will be added, enabling researchers to upload their own corpus and metadata, supporting multiple formats. Another extension concerns adding options for data analysis, and creating possibilities to sort, group, and filter search results and metadata.

Acknowledgements

The work on GrETEL was carried out in the framework of the following projects:

- Nederbooms: Exploitation of Dutch treebanks for research in linguistics (2010–2012) Flemish government, Department of Economy, Science and Innovation.
- Complement Raising and Cluster Formation in Dutch. A Treebank-supported Investigation (2011–2015) FWO (G.0.559.11.N.10).
- GrETEL2.0 (2013–2014) Dutch Language Union.
- AfriBooms (2013–2014) Dutch Language Union and Department of Arts and Culture of the Government of South Africa.
- CLARIN Educational Module GrETEL (2014) CLARIN-NL 14-008.
- SCATE (2014–2018) IWT SBO-130041.

[13] http://taalportaal.org

References

Liesbeth Augustinus and Frank Van Eynde. 2012. A Treebank-based Investigation of IPP-triggering Verbs in Dutch. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 7–12, Lisbon. Edições Colibri.

Liesbeth Augustinus and Frank Van Eynde. 2014. Looking for Cluster Creepers in Dutch Treebanks. Dat we ons daar nog kunnen mee bezig houden. *Computational Linguistics in the Netherlands Journal*, 4:149–170.

Liesbeth Augustinus and Frank Van Eynde. 2017. A Usage-based Typology of Dutch and German IPP verbs. *Leuvense Bijdragen*, 101:101–122.

Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-Based Treebank Querying. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3161–3167, Istanbul.

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2013. Example-Based Treebank Querying with GrETEL - now also for Spoken Dutch. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 423–428, Oslo. NEALT Proceedings Series 16.

Liesbeth Augustinus, Peter Dirix, Daniel van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard van Huyssteen. 2016a. AfriBooms: An Online Treebank for Afrikaans. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 677–682, Portorož.

Liesbeth Augustinus, Vincent Vandeghinste, and Tom Vanallemeersch. 2016b. Poly-GrETEL: Cross-Lingual Example-based Querying of Syntactic Constructions. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3549–3554, Portorož.

Liesbeth Augustinus. 2015. *Complement Raising and Cluster Formation in Dutch. A Treebank-supported Investigation*. LOT Dissertation Series 413. LOT, Utrecht.

Sjef Barbiers and Ineke Schuurman. 2015. Combined Case Study MIMORE - GrETEL. `http://www.meertens.knaw.nl/mimore/educational_module/case_study_mimore_gretel.html`.

Gosse Bouma, Marjo van Koppen, Frank Landsbergen, Jan Odijk, Ton van der Wouden, and Matje van de Camp. 2015. Enriching a Descriptive Grammar with Treebank Queries. In *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 13–25, Warsaw.

Peter-Arno Coppen. 2010. Bericht van de innerlijke stem. Synchronie en diachronie van de *heb-zoiets-van*-constructie. *Nederlandse Taalkunde*, 15(1):33–53.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454, Genoa.

Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Axel-Cyrille Ngonga Ngomo. 2010. The TIGER Corpus Navigator. In *Proceedings of the The 9th Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 91–102, Tartu.

Eric Hoekstra. 2010. *Van* als markeerder van zinnen in de directe en indirecte rede in het Fries en het Nederlands. *Leuvense Bijdragen*, 96:169–188.

Dan Klein and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Cambridge, MA. MIT Press.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket.

Gideon Kotzé, Vincent Vandeghinste, Scott Martens, and Jörg Tiedemann. 2016. Large Aligned Treebanks for Syntax-based Machine Translation. *Language Resources and Evaluation*. Vol. 51: 249–282.

Frank Landsbergen, Carole Tiberius, and Roderik Dernison. 2014. Taalportaal: an online grammar of Dutch and Frisian. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2206–2210, Reykjavik.

Scott Martens. 2013. TüNDRA: A Web Application for Treebank Search and Visualisation. In *Proceedings of the 12th International Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144, Sofia.

Paul Meurer. 2012. INESS-Search: A search system for LFG (and other) treebanks. In *Proceedings of the LFG'12 Conference. LFG Online Proceedings*, pages 404–421, Stanford.

Jan Odijk. 2015. Linguistic Research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN programme*, pages 219–247. Springer.

Philip Resnik and Aaron Elkiss. 2005. The Linguist's Search Engine: An Overview. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 33–36, Ann Arbor.

Ton van der Wouden, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman. 2002. Syntactic Analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 768–773, Las Palmas.

Ton van der Wouden, Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, and Jan Odijk. 2015. Enriching a grammatical database with intelligent links to linguistic resources. In *CLARIN Annual Conference 2015 Book of Abstracts*, pages 89–92, Wroclaw.

Frank Van Eynde, Liesbeth Augustinus, Ineke Schuurman, and Vincent Vandeghinste. 2014. Het verrassende resultaat van een copulativiteitspeiling. In Freek Van de Velde, Hans Smessaert, Frank Van Eynde, and Sara Verbrugge, editors, *Patroon en Argument. Een dubbelfeestbundel bij het emeritaat van William Van Belle en Joop van der Horst*, pages 47–62. Universitaire Pers, Leuven.

Frank Van Eynde, Liesbeth Augustinus, Ineke Schuurman, and Vincent Vandeghinste. 2016a. *Hebben* of *zijn* bij IPP's. *Leuvense Bijdragen*, 99-100:11–28.

Frank Van Eynde, Liesbeth Augustinus, and Vincent Vandeghinste. 2016b. Number agreement in copular constructions. A treebank-based investigation. *Lingua*, 178:104–126.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN programme*, pages 147–164. Springer.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In *Proceedings of TALN*, pages 20–42.

Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Joris Pelemans, Geert Heyman, Iulianna van der Lek-Ciudin, Arda Tezcan, Donald Degraen, Jan Van den Bergh, Lieve Macken, Els Lefever, Marie-Francine Moens, Patrick Wambacq, Frieda Steurs, Karin Coninx, and Frank Van Eynde. 2016. Smart Computer Aided Translation Environment. In *Baltic Journal of Modern Computing. Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, volume 4 (2), page 382, Riga.

Ventislav Zhechev. 2009. *Automatic Generation of Parallel Treebanks: An Efficient Unsupervised System*. Dublin City University.