GrETEL
Searching for breadcrumbs in texts

(CLARIN Educational Module)

Liesbeth Augustinus
Ineke Schuurman
Vincent Vandeghinste
Frank Van Eynde
{liesbeth,ineke,vincent,frank}@ccl.kuleuven.be

Centre for Computational Linguistics
KU Leuven

October, 2014

Contents

1 Treebanks and Linguistics

1.1 Introduction

The last decades the use of text corpora containing authentic language samples has been of major importance in the study of language. Within theoretical linguistics, corpus examples may provide empirical evidence for the linguistic constructions under investigation, thereby motivating or refuting certain assumptions made by the theory. Besides being a source of linguistic constructions, corpora also allow a quantitative investigation of language, providing insight into absolute and relative frequencies of constructions.

Corpora come in several types, sizes and formats. For instance, one can differentiate synchronic from diachronic corpora, monolingual from multilingual corpora, parallel from not-parallel corpora, plain text corpora from annotated corpora, etc.

The kind of corpora that are used within corpus linguistics is largely dependent on the subject of the research and the availability of the corpora. For some languages, no corpora or only plain text ones are available. Extracting information from raw data is very labour intensive, especially if one is looking for non-lexical phenomena. The use of annotated data makes such tasks easier, provided that one is familiar with the annotation guidelines. The level of data annotation ranges from annotations on the lexical level (e.g. lemmatization, morphology and part-of-speech tagging) to annotations on the syntactic level (e.g. dependency relations, syntactic categories) and the semantic level (e.g. semantic roles), as well as annotations on discourse level.

For syntactic research, the most obvious choice is to use syntactically annotated corpora, also known as *treebanks*. For Dutch, several treebanks are available, such as the Alpino Treebank (van der Beek et al., 2002), LASSY Small and Lassy Large (van Noord et al., 2013), CGN (Oostdijk et al., 2002), and SoNaR (Oostdijk et al., 2013).

Some treebanks are manually constructed, such as the CGN treebank, while others are automatically created, such as SoNaR. In the latter case, a *parser* is used to add syntactic information to each sentence in the corpus. A commonly used parser for Dutch is the Alpino parser (van Noord, 2006). It outputs trees that have a similar structure to the trees in the treebanks mentioned above. Figure 1 shows an example of a parse tree for the sentence *De man eet een appel* ‘The man is eating an apple.’

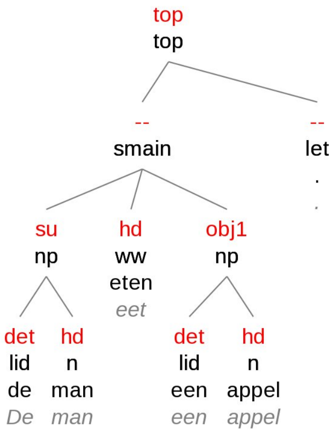


Figure 1: Syntax tree created with the Alpino parser

The tree contains linguistic annotations at various levels. On the word level, it contains lemmas and labels for word class or part-of-speech (POS), i.e. the kind of information that is included in many ‘flat’, i.e., raw or POS-tagged, corpora as well. Besides annotation on the word level, the parse also includes constituent structure (e.g. the noun phrase NP *de man*) and dependency information (e.g. the subject SU *de man*, the direct object OBJ1 *een appel*).

1.2 Why Treebanks?

Similar to other corpora, treebanks can be used for both qualitative and quantitative linguistics. For example, one can use a treebank to explore a linguistic construction, or to find real language data in order to support or refute a theoretical claim. The set of relevant examples may furthermore provide information on the context in which the construction under investigation occurs. This makes it possible to investigate which parameters are related to the phenomenon under investigation, which may be hard to determine by means of introspection.
In addition, one can extract frequency information from treebanks, which provides quantitative information about the constructions under investigation.

Treebanks are especially interesting for syntactic research, as the phenomena that are investigated typically generalize over both word forms and word order. Those phenomena are usually hard to extract exhaustively from flat corpora, as these do not contain information beyond the word level. As it is possible to generalize over lexical patterns, it is not necessary to spell out complete paradigms in the search instruction. As a result, less examples will be missed out.

Moreover, as treebanks contain more (detailed) annotation than flat corpora, it is possible to define more precise queries, which reduces the noise in the search results (they return less *false positives*).

Even though the use of syntactic annotations reduces the complexity of search instructions if one is looking for non-lexical patterns, the annotation also introduces some complexity in comparison to searching in flat corpora. For example, one has to be familiar with the annotations (e.g. one has to know the grammar used in the treebank (e.g. the Dutch treebanks that will be introduced here do not have unary branching NPs). However, some search tools try to compensate for this drawback, e.g. the Linguist's Search Engine

(LSE) (Resnik & Elkiss, 2005)[1], and GrETEL (Augustinus et al., 2012Augustinus et al., 2013).

As mentioned before, for Dutch, several treebanks are available. Three of them are contained in GrETEL:

The CGN Treebank The Corpus Gesproken Nederlands (CGN) (Oostdijk et al., 2002) is an annotated corpus of spoken Dutch.[2] It consists of recorded speech which is orthographically transcribed, resulting in a corpus of ca. 10 million words, of which 1 million is syntactically analysed (and manually verified). That syntactically annotated part of CGN is referred to as the *CGN treebank*.

Two thirds of the corpus data consists of Dutch as spoken in the Netherlands (NL-Dutch), whereas one third of the data comprises Dutch as spoken in Flanders[3] (VL-Dutch). The corpus contains dialogues as well as monologues, and is further divided into specific genres (e.g. telephone conversations, lectures). The division into subcorpora allows to investigate stylistic variation (e.g. by comparing spontaneous conversations to news reports), as well as regional variation (by comparing VL-Dutch to NL-Dutch).

LASSY Small The LASSY treebank (Large Scale Syntactic Annotation of written Dutch) (van Noord et al., 2013) is a corpus of syntactically annotated sentences.[4] The LASSY project resulted in the construction of two treebanks: LASSY Small and LASSY Large. Both treebanks were parsed using the Alpino parser (van Noord, 2006). LASSY Small is manually corrected, whereas the LASSY Large is not. LASSY Small is complementary to the CGN treebank. As the corpora are more or less equal in size, they are suited for comparing written to spoken language data. Moreover, the annotations used for both tree-

[1] Unfortunately, the LSE is no longer running.

[2] `http://lands.let.ru.nl/cgn`

[3] The Dutch speaking part of Belgium.

[4] `http://www.let.rug.nl/~vannoord/Lassy`

banks are very similar, so one can query both treebanks in GrETEL (cf. section 2) using the same queries. It is not possible to compare VL-Dutch versus NL-Dutch in LASSY,[5] but one can compare different genres (e.g. newspaper text versus law texts).

SoNaR The SoNaR corpus (Oostdijk et al., 2013) is a balanced corpus of written Dutch that consists of 500 million words. It is automatically tokenized, POS-tagged, lemmatized, and syntactically analysed, using the Alpino parser (van Noord, 2006). In contrast to the CGN and LASSY treebanks, it is not manually corrected, but its large size makes it suitable to investigate infrequent phenomena.

1.3 Search tools

There exist several search tools to query treebanks. The following search tools can be used for the Dutch treebanks mentioned above:

TigerSearch (Lezius, 2002) is a graphical user interface for querying treebanks encoded in the TIGER-XML data format, like CGN.[6] The data can be queried using the TIGER language.[7]

Dact (van Noord et al., 2013)[8] is a graphical user interface to query treebanks encoded in the Alpino-XML data format. Treebanks included in Dact can be queried using the XPath query language, which is a W3C standard for querying XML trees.[9]

[5] In the remainder of this paper, LASSY refers to LASSY Small only.

[6] The version of the CGN treebank used in GrETEL is converted in another XML-format (Alpino-XML), cf. http://www.let.rug.nl/vannoord/Lassy/alpino_ds.dtd

[7] For a manual on both TIGERSearch and the TIGER query language, see http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/manual_html.html.

[8] http://rug-compling.github.io/dact

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

GrETEL (Augustinus et al., 2012Augustinus et al., 2013) is a linguistic search engine for Dutch treebanks. It provides two ways to query treebanks: via a natural language example and via an XPath query, see section 2.

1.4 Examples of treebank-based syntactic research

That treebanks are a valuable resource for linguistics has been shown in several studies on a variety of linguistic topics. For instance, van der Beek (2005) has used different Dutch corpora and treebanks to investigate cleft constructions, the dative alternation and determinerless PPS. Van Eynde (2009) has used the CGN treebank to investigate copular constructions. Augustinus & Van Eynde (2012) have used treebanks to empirically study the IPP effect in Dutch. Augustinus & Eynde (2015) provide a treebank account of Dutch verb clusters and verb cluster interruption.

1.5 Primary literature

- Meurers, D. & S. Müller. 2009. Corpora and Syntax. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, chap. 42, 920–933. Berlin: Mouton de Gruyter

- van Noord, G., G. Bouma, F. Van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang & V. Vandeghinste. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In P. Spyns & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: resources, tools and applications*, Springer

- Van Eynde, F., L. Augustinus, I. Schuurman & V. Vandeghinste. 2014. Het verrassende resultaat van een copulativiteitspeiling. In F. Van de Velde, H. Smessaert, F. Van Eynde & S. Verbrugge (eds.), *Patroon en argument. een dubbelfeestbundel bij het emeritaat van william van belle en joop van der horst*, 47–62. Leuven: Universitaire Pers

- Augustinus, L. & F. Van Eynde. 2015. Looking for Cluster Creepers in Dutch Treebanks. *Dat we ons daar nog kunnen mee bezig houden. Computational Linguistics in the Netherlands Journal* 4

1.6 Secondary literature

- Abeillé, A. (ed.). 2003. *Treebanks. Building and Using Parsed Corpora*. Dordrecht: Kluwer

- Bouma, G. & G. Kloosterman. 2002. Querying dependency treebanks in XML. In *Proceedings of the Third international Conference on Language Resources and Evaluation (LREC-2002)*, Gran Canaria

- Bouma, G. & G. Kloosterman. 2007. Mining Syntactically Annotated Corpora using XQuery. In *Proceedings of the Linguistic Annotation Workshop (ACL 07)*,

- Nivre, J. 2008. Treebanks. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 1, chap. 13, 225–241. Berlin: Mouton de Gruyter

- The proceedings of the *International Workshop on Treebanks and Linguistic Theories* (TLT). This conference provides a forum for researchers working on building treebanks and treebanking tools as well as the exploration of treebanks for linguistic purposes.

2 GrETEL

2.1 Introduction

GrETEL (**Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics) is a search engine for treebanks. It is the result of the Nederbooms project, a CLARIN[10]

[10] `http://www.clarin.eu/`

project which aimed at the development of user-friendly tools for the exploitation of treebanks by linguists who are not familiar with language technology. The most recent version was created in the GrETEL 2.0 project.

GrETEL is accessible online,[11] implicating that users do not have to install any treebanks or specific software (e.g. a parser) locally.[12] It has two search modes: *Example-based search* and *XPath search*, which are presented in section 2.2 and 2.3. GrETEL is used for the treebank-based linguistic case studies in the remainder of this module.

2.2 Example-based search

(Descriptive) linguists are often reluctant towards using treebanks because of, on the one hand, the limited user-friendliness of the query languages and search tools, and on the other hand, the lack of standardisation in treebanks and query languages. In order to compensate for those problems, GrETEL provides the example-based querying option, starting from a natural language example instead of a formal search instruction.

Below, the successive steps in the process are described on the basis of an example.

Step 1: Provide an example The GrETEL search engine does not ask for any formal input query. The tool takes something linguists are familiar with as input: natural language. Since linguists tend to start their research from example sentences, the methodology of example-based querying allows users to search for similar constructions as the example they provide. How similar is up to the user to decide.
For example, if one is looking for constructions in which a verb appears as

[11] `http://gretel.ccl.kuleuven.be`

[12] The user is adviced to access the tools via Mozilla Firefox (`http://www.mozilla.org`).

an infinitive if it is selected by an auxiliary of the perfect,[13] one could feed example (1) to the system. Such constructions are diffcult to find in a flat corpus.

(1) dat hij Marie heeft horen zingen.
that he Marie has hear.IPP sing.INF
'that he has heard Marie sing.'

In the example-based search mode, the user can choose between *basic* and *advanced* search. The advanced search mode has more search options compared to the basic search mode, and it is possible to adapt the search instruction (the query). In this example, the basic search mode is chosen.[14]

Step 2: The parse tree GrETEL parses the input example with the Alpino parser and returns the input as a syntax tree. The search instruction will be based on this parse tree, so if the syntactic analysis of the example is different from the intended analysis, the user is adviced to choose another input example.

Figure 2 shows a correct parse of the input construction in (1), so the example can be used for querying the treebanks.

Step 3: The selection matrix In this step, the user can indicate for each word whether (s)he is interested in the word class, the lemma, or the (exact) word form (or token). To indicate that words are not part of the target construction, one can use the OPTIONAL nodes button (cf. Figure 3). Note that both the dependency relation (`rel`) and the phrasal category (`cat`) of all relevant nodes are taken into account.

Since we are looking for IPP constructions, the lemma of the perfect auxiliary is relevant, as well as the POS of the infinitives. The other nodes are not relevant for retrieving the target construction.

[13] Known as Infinitivus Pro Participio, or IPP, cf. chapter 5.1 for a thorough analysis.

[14] How in advanced mode the query can be adapted in order to broaden the query is explained at page **??**.

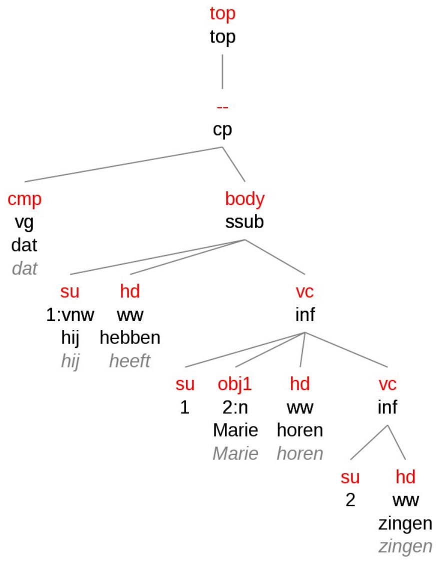
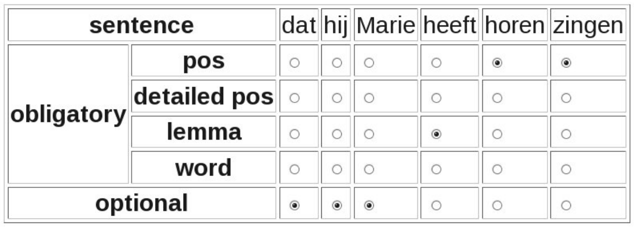


Figure 2: Parsed version of the input example



sentence		dat	hij	Marie	heeft	horen	zingen
obligatory	pos	○	○	○	○	◉	◉
	detailed pos	○	○	○	○	○	○
	lemma	○	○	○	◉	○	○
	word	○	○	○	○	○	○
optional		◉	◉	◉	○	○	○

Figure 3: Selection matrix

Step 4: Selection of a treebank The user can choose which treebank (s)he wants to query. One can choose between the CGN treebank, the LASSY treebank, and SoNaR. For CGN and LASSY, it is possible to query the treebank as a whole, or to query only certain components of a treebank (for instance, only the VL-Dutch parts of CGN). For SoNaR, it is only possible to select one component at a time. In this example, the complete CGN treebank is searched.

Step 5: Query overview GrETEL extracts the information provided in the selection matrix from the parse tree (Figure 2), which results in the query tree in Figure 4.

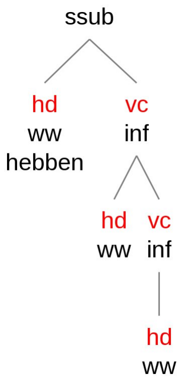


Figure 4: Query tree based on the input construction

The query tree is automatically converted into an XPath expression (2), used to query the treebank. In the basic search mode, the query is not shown to the user at this stage. In the advanced search mode,[15] however, users can optionally adapt the XPath expression in order to refine or generalize the search instruction.[16]

(2)
```
//node[@cat="ssub" and node[@rel="hd" and @pt="ww" and
@lemma="hebben"] and node[@rel="vc" and @cat="inf" and
node[@rel="hd" and @pt="ww"] and node[@rel="vc" and
@cat="inf" and node[@rel="hd" and @pt="ww"]]]]
```

Step 6: Results In the last phase the search results are presented, i.e. the sentences containing the construction at hand. The user can inspect the tree and/or or the source XML of the results. It is also possible to download the

[15] See also page 16ff.

[16] Using the XPath expression in (2) the results are limited to verb-final constructions (SSUB), and constructions in which the finite verb selects bare infinitives. The query thus does not take into account IPP constructions with *te*-infinitives (TI), or verb-initial constructions (SMAIN and SV1).

results in text format. For the query in (2), 79 matches in 76 sentences were found in the CGN Treebank. Some examples are given in (3).

(3) a. dus uh ik vermoed dus dat de journalisten de film twee keer
thus uh I assume thus that the journalists the movie two times
hadden moeten zien.
had must.IPP see.INF
'so uh I assume that the journalists had to watch the movie twice.' [CGN, fvf600243__10]

b. dus 't is niet dat we daar echt mensen hebben leren kennen.
so it is not that we there really people have learn.IPP know.INF
'so it is not the case that we really have learned to know people over there.' [CGN, fvb400155__296]

Note that the 'Gr' in GrETEL stands for *greedy search*. This means that the matches may include constructions in which nodes appear between the nodes defined in the query tree. An example is the separable verb particle *mee* 'with' in (4), which appears within the IPP construction:

(4) ...dat Joegoslavië eigenlijk uh niet had mee mogen doen.
...that Yugoslavia actually uhm not had with may.IPP do
'...that Yugoslavia in fact was not allowed to join.' [CGN, fnl007393__74]

Using the advanced mode, users can adapt the XPath expressions used for search.

For example, if one wants to include verb-initial sentences,[17] constructions with the auxiliary *zijn* 'be', and constructions in which the IPP verb selects a *te* infinitive or a terminal VC node as well, one could – in the advanced search mode – generalize the query in (2) to the query in (5):

(5)
```
//node[@cat and node[@rel="hd" and @pt="ww" and
(@lemma="hebben" or @lemma="zijn")] and node[@rel="vc"
and @cat="inf" and node[@rel="hd" and @pt="ww"]
and node[@rel="vc" and (@cat="inf" or @cat="ti" or
@pt="ww")]]]
```

[17] Cf, page 15

By underspecifying the `@cat` feature for the top node, verb-initial as well as verb-final sentences will be included in the results. The `or` operator is used to include both *hebben* and *zijn*, and to extend the VC complement types. A terminal VC has the `@pt="ww"` feature and a *te* infinitive contains the `@cat="ti"` feature.

The query in (5) returns 792 matches in 777 sentences in CGN. Some examples are presented in (6). The verb-initial sentences clearly show the *greedy* search, since in those clauses the finite verb and the infinitives are usually not adjacent. Such constructions would be hard to extract in an efficient way using a flat corpus.

(6) a. de organisatie heeft een groot aantal partytenten uit België
the organisation has a big number party-tents from Belgium
laten komen.
let.IPP come.INF

'the organisation has got a large number of party tents coming from Belgium.' [CGN, fnk005386__3]

b. ik heb m'n nicht proberen te bellen want die uh ...
I have my cousin try.IPP to call.INF because that one uh

'I have tried to call my cousin because she erm ...' [CGN, fna000628__73]

If one does not want to tinker with the XPath query, one can also build separate queries using slightly different input examples.

2.3 XPath Search

Besides querying treebanks by example, it is also possible to query the CGN and LASSY treebanks by means of an XPath expression straightaway, using the *XPath Search* version of GrETEL. Similar to Dact, users have to provide the XPath queries themselves.[18]

[18] http://www.xpath.org

3 Some basic use cases

Although the GrETEL search engine is *in se* language independent, the version discussed in this module[19] is devoted to Dutch corpora, therefore at least a basic knowledge of Dutch is necessary to do the assignments, and to understand the use cases presented in the following chapters.
Familiarize yourself with GrETELTry through the following exercises:

Assignment 1: Look for main clauses with *hun* ('them') as indirect object in CGN. Use the basic search mode.
Do the same for constructions with *hen* ('them') as indirect object.

Assignment 2: Look for sentences in LASSY (using the basic search mode) with a constituent (not a modifier) expressing a quantity, e.g. *de vergadering duurde drie uur* 'the meeting lasted three hours' (vs *hij heeft drie uur naar de tv gekeken* 'he watched TV for three hours'.) Give a list of the verbs selecting such arguments. Compare your findings with what is stated in Haeseryn et al. (1997), does it contain a list with such verbs? If not, try to find such a list on the Internet.

***

In the next two chapters some case studies will be presented, one of binominal constructions and one on IPP (Infinitivus Pro Participio).

[19] There is also a version for a corpus in Afrikaans (http://gretel.ccl.kuleuven.be/afribooms), and a version for other languages could be made.

4 GrETEL case study I: Binominal constructions containing a quantificational noun

4.1 Introduction

QUANTIFICATIONAL NOUNS are nouns serving to quantify something, be it sometimes in an imprecise way. In the sentences (7) and (8), *number* and *deal* are examples of such nouns ((Huddleston & Pullum, 2002, 339))

(7) A number of problems remain

(8) He wastes a great deal of time

Constructions of type "*een aantal mensen*" (a number of people), i.e. NPs of type '[indef-ART] [quantificational noun-SG] [common noun-PL]' functioning as the subject of a sentence are a conundrum in Dutch: Should the finite verb be in singular or plural form? The choice depends on what is considered to be the *central* element of this NP, i.e. the *head*, the first noun (a singular form) or the second one (plural).
The idea is that in principle in constructions with nouns expressing a vague (number, mass) or approximate (handful, (some/a) ten) quantity, both Ns can function as head of the NP.

4.2 According to the *ANS* ...

According to ANS (Algemene Nederlandse Spraakkunst), there are several types of binominal constructions (Haeseryn et al., 1997) in Dutch.

(9) een tiental, drietal, ... (approximately ten, approximately three, ...)

a. een tiental honden is/zijn ontsnapt
a ten dogs is/are escaped
'some ten dogs have escaped'

b. het/dit tiental honden is/*zijn ontsnapt
the/this ten dogs is/*are escaped
’some/these nine or ten dogs have escaped’

(10) een aantal/handvol/massa (a number/handful/mass/...)

a. een aantal reizigers klaagt/klagen
a number passenger complain.SG/complain.PL
‘a number of passengers are complaining’

b. een handvol piloten staakte/staakten
a handful pilots striked
‘a handful of pilots have been on strike’

c. het/dit aantal reizigers daalt/*dalen
the number passengers drop.SG/*drop.pl
’the number of passengers is dropping’

(11) een/het/dit soort (a/the/this kind)

a. dit soort problemen ontstaat/ontstaan geregeld /klagen
this kind problems arise.SG/arise.PL
‘this kind of problems often arises’

(12) een vlucht/zwerm/rij/groep/... (a flight/swarm/row/group/...)

a. een/de/die vlucht regenwulpen streek/*streken neer
a/the/this flight whimbrels settled.SG/settled.PL on
‘a flight of whimbrels were settling’

b. een/het/dat groepje toeristen stond/*stonden voor de Nachtwacht
a/the/that group-DIM tourists stand.SG/stand.PL for the Nachtwacht
‘a/the/that small group of tourists was standing in front of the Nachtwacht’

(Haeseryn et al., 1997) claim that the type shown in (9) and (10) can be combined with both plural and singular finite verbs, although for plural forms the article should be indefinite. A definite article will make the construction

	indef.art.	**def. art.**	DEM. PRON.
type *tiental, aantal*	sg, pl	sg	sg
type *soort*	sg, pl	sg, pl	sg, pl
type *vlucht*	sg	sg	sg

Table 1: Plural of singular finite verb?

ungrammatical, cf example 10c.[20] Constructions of type (11) allow for plural finite verbs, even when the determiner is a definite article or a demonstrative pronoun,[21] whereas constructions of type (12) will never occur with a plural finite verb when the first noun is a singular one. All possibilities are shown in Table 1. In this section occurrences with demonstratice pronouns are largely ignored.

With respect to constructions of type (10), (Haeseryn et al., 1997) mention that especially *aantal, hand(je)vol, hoop, massa* (number, handful, lot, mass) and *reeks* (series) are used in such constructions.
Is all of this reflected in *real* language use? Or can counterexamples be found in written or spoken corpora?

In the following, the focus will be on constructions in Lassy and CGN containing

(A) a determiner plus a series of ar least two common nouns, differing in number (sg vs pl), functioning as subject

(B) a determiner plus a series of at least two common nouns (same number), functioning as subject, combined with a finite verb showing a different number

[20] Note that a demonstrative pronoun, such as *dit* (dit), has the same effect.

[21] Only for this type, where combinations with such a pronoun are explicitly mentioned in (Haeseryn et al., 1997), the occurence of a demonstrative pronoun is queried. For the other constructions, just articles were taken in consideration.

The hypotheses are:

1. the list in ANS is not comprehensive wrt the type of the first N
2. real 'counterexamples', if any, will be found in spoken, i.e., more spontaneous, language.

4.3 Binominal constructions in Lassy and CGN

GrETEL is used in the ADVANCED SEARCH mode to find sentences with the patterns described above. This ADVANCED mode enables optimization of the XPath expression used to query the corpora.[22]

The input sentence:
een aantal reizigers klaagt (a number of passengers are complaining)

This sentence contains the pattern INDEF-NS-NP-SG (an indefinite article, followed by a singular common noun, followed by a plural common noun in combination with a finite verb (singular)). It is analysed as shown in figure (5).[23]
Indicating as relevant for the nouns[24] and the verb the DETAILED WORD CLASS and for the article the LEMMA, and ignoring the properties of the dominating node, cf. figure 6, the query tree in figure 7 and the XPath expression in (13) were generated:

```
(13)  //node[@cat and node[@rel="su" and @cat="np" and
      node[@rel="det" and @pt="lid" and @lemma="een"] and
```

[22] As neither the corpora used (Lassy and CGN) nor the tagger and parser (Alpino) differentiate between quantificational common nouns and other common nouns, the search will take into account all subject NPs containg two or more common nouns.

[23] Note that only 'real' nouns are taken into account, nominalized adjectives etc (as in "*een aantal bejaarden*" (a number of elderly people) are neglected.

[24] In this chapter only 'real' nouns are taken into account, i.e., nominalized adjectives (like *jarige* in *de jarige moet trakteren* or nominalized verbs (like *gepensioneerde* in *de gepensioneerde heeft makkelijk praten* are ignored.

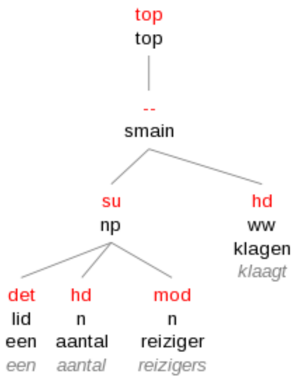


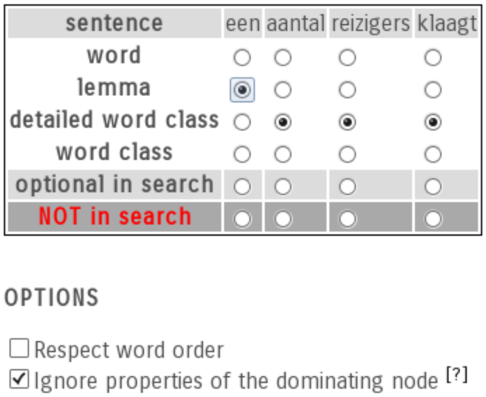
Figure 5: Syntax tree (Alpino) for ’een aantal reizigers klaagt’

```
node[@rel="hd" and @pt="n" and @graad="basis" and
@genus="onz" and @getal="ev" and @naamval="stan" and
@ntype="soort"] and node[@rel="mod" and @pt="n" and
@getal="mv" and @graad="basis" and @ntype="soort"]]
and node[@rel="hd" and @pt="ww" and @pvtijd="tgw"
and @wvorm="pv" and @pvagr="met-t"]]
```

This XPath was adapted in order a) to search for neuter as well as non-neuter nouns, and b) to search for past as well as present tense, and for all (singular) forms of agreement:

(14)
```
//node[@cat and node[@rel="su" and @cat="np" and
node[@rel="det" and @pt="lid" and @lemma="een"] and
node[@rel="hd" and @pt="n" and @graad="basis" and
@getal="ev" and @naamval="stan" and @ntype="soort"]
and node[@rel="mod" and @pt="n" and @getal="mv"
and @graad="basis" and @ntype="soort"]] and
node[@rel="hd" and @pt="ww" and @wvorm="pv" and
(@pvagr="met-t" or @pvagr="ev")]]
```

For Lassy, this results in 57 hits. And, surprisingly (?), in only 10 hits for CGN.



sentence	een	aantal	reizigers	klaagt
word	○	○	○	○
lemma	●	○	○	○
detailed word class	○	●	●	●
word class	○	○	○	○
optional in search	○	○	○	○
NOT in search	○	○	○	○

OPTIONS

☐ Respect word order
☑ Ignore properties of the dominating node [?]

Figure 6: The matrix for ’een aantal reizigers klaagt’

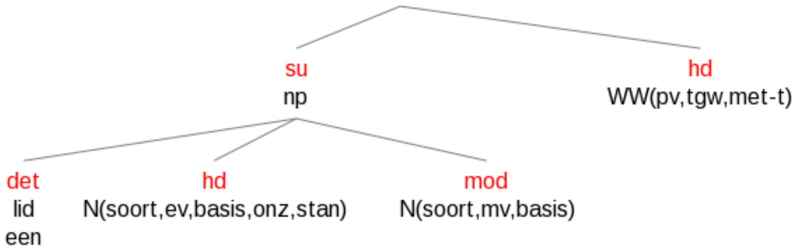


Figure 7: Search tree for ’een aantal reizigers klaagt’

In the remainder of this section, the abbreviations shown in Table 2 are used.[25] Note that in principle, sentences with the last noun (often the second one) as head do not come with a flat structure in Alpino![26]

(15) a. een aantal reizigers klagen
a number passengers complain.PL

[25] In these tables zero hits are represented by a ’-’ (dash).

[26] But, as shown in Table 3, there are several instances in both treebanks in which according to the number of the finite verb, this verb agrees with the second noun instead of the first one, and still a flat analysis was used.

def	definite article
indef	indefinite article
N	common noun, functioning as head of the subject NP
n	non-head common noun
s	singular noun
p	plural noun
sg	finite verb, singular
pl	finite verb, plural

Table 2: abbreviations used

	Lassy	CGN	**total**		Lassy	CGN	**total**	**TOTAL**
indef-Ns-ns-sg	14	8	**22**	def-Ns-ns-sg	19	1	**20**	**42**
indef-Ns-ns-pl	-	-	-	def-Ns-ns-pl	-	1	**1**	**1**
indef-Ns-np-sg	57	10	**67**	def-Ns-np-sg	73	12	**85**	**152**
indef-Ns-np-pl	17	-	**17**	def-Ns-np-pl	4	-	**4**	**21**
indef-Np-np-sg	-	-	**-**	def-Np-np-sg	-	-	**-**	**-**
indef-Np-np-pl	-	-	**-**	def-Np-np-pl	2	-	**2**	**2**
indef-Np-ns-sg	-	-	**-**	def-Np-ns-sg	1	-	**1**	**1**
indef-Np-ns-pl	-	-	**-**	def-Np-ns-pl	1		**1**	**1**
total	**88**	**18**	**106**		**100**	**14**	**114**	**220**

Table 3: flat NPs

‘a number of passengers are complaining’

b. een aantal reizigers klagen over de vertragingen
a number passengers complain.PL about the delays
‘a number of passengers are complaining about the delays’

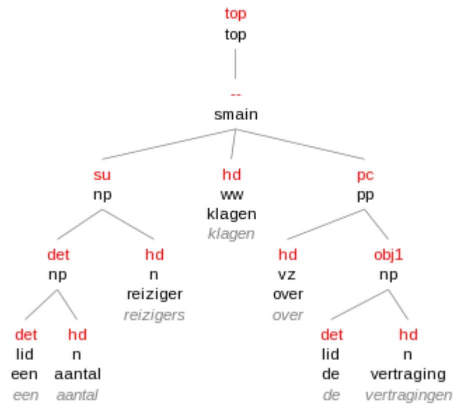


Figure 8: Correct tree (Alpino) parser for ’een aantal reizigers klagen ...’

This results in a Xpath query as in (16)

(16)
```
//node[@cat and node[@rel="su" and @cat="np" and
node[@rel="det" and @cat="np" and node[@rel="det"
and @pt="lid" and @lwtype="onbep" and @npagr="agr"
and @naamval="stan"] and node[@rel="hd" and
@pt="n" and @graad="basis" and @genus="zijd" and
@getal="ev" and @naamval="stan" and @ntype="soort"]]
and node[@rel="hd" and @pt="n" and @getal="mv"
and @graad="basis" and @ntype="soort"]] and
node[@rel="hd" and @pt="ww" and @pvtijd="tgw" and
@wvorm="pv" and @pvagr="mv"]]
```

which is modified in order to become a more general version

	Lassy	CGN	**total**		Lassy	CGN	**total**	**TOTAL**
indef-ns-Ns-sg	4	15	**19**	def-ns-Ns-sg	-	-	**-**	**19**
indef-ns-Ns-pl	-	-	-	def-ns-Ns-pl	-	-	**-**	**-**
indef-ns-Np-sg	1	15	**16**	def-ns-Np-sg	-	6	**6**	**22**
indef-ns-Np-pl	46	73	**119**	def-ns-Np-pl	-	-	**-**	**119**
indef-np-Np-sg	-	-	**-**	def-np-Np-sg	-	-	**-**	**-**
indef-np-Np-pl	-	-	**-**	def-np-Np-pl	-	-	**-**	**-**
total	**51**	**103**	**154**		**-**	**6**	**6**	**160**

Table 4: structured NPs

(17)
```
//node[@cat and node[@rel="su" and @cat="np" and
node[@rel="det" and @cat="np" and node[@rel="det"
and @pt="lid" and @lwtype="onbep"] and
node[@rel="hd" and @pt="n" and @graad="basis" and
@getal="ev" and @naamval="stan" and @ntype="soort"]]
and node[@rel="hd" and @pt="n" and @getal="ev"
and @graad="basis" and @ntype="soort"]] and
node[@rel="hd" and @pt="ww" and @wvorm="pv"]]
```

The numbers mentioned in Table 3 and Table 4 are lower than the total number of hits when using the XPath queries as mentioned above, not all results being of the type we are looking for.
Sometimes this is due to mistakes in the corpora, either spelling (splitting words instead of writing them as one word: *rijst verbouw* (rice cultivation) instead of *rijstverbouw*) or tagging mistakes: *een*-Det *wijze*-N *belegger*-N (a wise_person investor) instead of *een*-Det *wijze*-Adj *belegger*-N (a prudent investor). Other times the type of binominal construction is not the one we are looking for:

(18) *leraar geschiedenis* (teacher of history), *directeur gemeentewerken* (director of the public works depertment), *testosteron substitutietherapie* (testoteron substitution therapy)

(19) *het domeinnaamsysteem (DNS)* (the domain name system (DNS)) , *het gen (DNA)* (the gene (DNA))

In such constructions the number of the head (the first noun) will be the same as that of the finite verb. These constructions are not taken into consideration in tables 3 and 4.

4.4 A closer look

FIRST NOUN	FLAT	FLAT	STRUCT.	STRUCT.	TOTAL
	Lassy	CGN	Lassy	CGN	
aandeel (part)	3	-	-	-	3
aantal (number, amount)	87	14	25	49	175
baal (bale)	1	-	-	-	1
bende (mass)	-	-	-	1	1
bos (bunch)	-	1	-	-	1
brigade (brigade, team)	2	-	-	-	2
bundel (bundle, wad)	1	-	-	-	1
bus (bus, tin)	-	1	-	-	1
categorie (category)	2	-	-	-	2
collectie (collection)	2	-	-	-	2
concentratie (concentration)	2	-	-	-	2
doos (box)	-	1	-	-	1
dosis (dose)	1	-	-	-	1
druppel (drop)	-	1	-	-	1
gedeelte (part)	-	1	-	-	1
gehalte (proportion)	2	-	-	-	2
generatie (generation)	4	-	-	-	4
glas (glass)	3	-	-	-	3
groep (group)	10	1	-	1	12
hand(je)vol (handful)	1	-	-	-	1
heleboel (lot)	-	-	2	8	10

continued on next page

FIRST NOUN	FLAT	FLAT	STRUCT.	STRUCT.	TOTAL
	LASSY	CGN	LASSY	CGN	
hoeveelheid (amount)	13	-	-	-	13
hoop (pile, deal)	-	-	1	7	8
kolonie (colony)	2	-	-	-	2
maximum (maximum)	3	-	-	-	3
menigte (crowd)	1	-	-	-	1
meter (metre)	-	1	-	-	1
miljard (billion)	3	-	-	-	3
miljoen (million)	13	-	2	-	15
minimum (minimum)	2	-	-	-	2
minuut (minute)	1	-	-	-	1
lading (load)	1	-	-	-	1
paar (couple)	-	-	13	25	38
pak (pack)	-	-	-	2	2
partij (set)	-	1	-	-	1
percentage (percentage)	2	-	-	-	2
ratio (proportion)	1	-	-	-	1
reeks (series)	2	2	3	-	7
rij (row)	1	1	-	-	2
schare (crowd)	1	-	-	-	1
serie (series)	2	-	-	-	2
som (sum)	1	-	-	-	1
soort (kind)	2	-	3	10	15
stel (pair)	-	-	-	1	1
strook (strip)	1	-	-	1	2
stroom (stream)	2	2	-	-	4
stuk (piece)	3	1	-	2	6
type (type)	1	-	-	-	1
verzameling (collection)	-	1	-	-	1
voorraad (supply)	1	-	-	-	1
weinig (little)	-	1	-	-	1
wolk (cloud, flock)	1	-	-	-	1

continued on next page

first noun	flat	flat	struct.	struct.	TOTAL
	Lassy	CGN	Lassy	CGN	
x-tal (number)	7	-	2	2	11
zak (pack)	1	1	-	-	
zwerm (swarm, flock)	1	-	-	-	1

Table 5: The first nouns encountered

As mentioned above, according to (Haeseryn et al., 1997), especially *aantal, hand(je)vol, hoop, massa* (number, handful, lot, mass) and *reeks* (series) are used in binominal constructions. For the moment ignoring the specific types, Table 5 shows that *aantal* (number) is used very often. The other ones (*hand(je)vol, hoop, reeks* (handful, lot, series)) a few times, except for *massa* (mass): it does not occur at all in these corpora, at least not in binominal quantificational constructions functioning as subject!

Looking at first nouns in a binominal construction that appear with singular as well as plural finite verbs:

first noun	L-ind-ns-Np-pl	C-ind-ns-Np-pl	L-ind-Ns-np-pl	TOTAL
aantal	25	33	4	62
bende	-	1	-	1
groep	-	1	1	2
heleboel	2	8	-	10
hoeveelheid	-	-	-	-
hoop	-	4	-	4
miljard	-	-	1	1
miljoen	2	-	-	2
paar	12	24	-	36
pak	-	2	-	2
reeks	3	-	1	4
rij	-	-	-	-
X-tal	2	1	2	4

Table 6: Indefinite - Singular first noun - plural finite verb

FIRST NOUN	L-DEF-NS-NP-PL	C-DEF-NS-NS-PL	TOTAL
aantal	-	1	1
bende	-	-	-
groep	-	-	-
heleboel	-	-	-
hoeveelheid	1	-	1
hoop	-	-	-
miljard	-	-	-
miljoen	1	-	1
paar	-	-	-
pak	-	-	-
reeks	-	-	-
rij	1	-	1
X-tal	-	-	-

Table 7: Definite - singular first noun - plural finite verb

The example in table 7 for *aantal* is in fact one with three nouns instead of two (GrETEL being greedy):

(20) en het totaal aantal dieren die bij de verschillende boeren
and the total number animals that by the different farmers
staan, moeten ...
stand, must ...
and the total amount of animals stalled by these farmers, are to be ...
[CGN, fvg600014__124]

We did not find any real examples of binominal constructions (subject position) where the number of the finite verb doesn't match withg that of at least one of the nouns involved.
First nouns (singular) combined with a definite article and a plural finite verb are:

- *hoeveelheid* (amount)

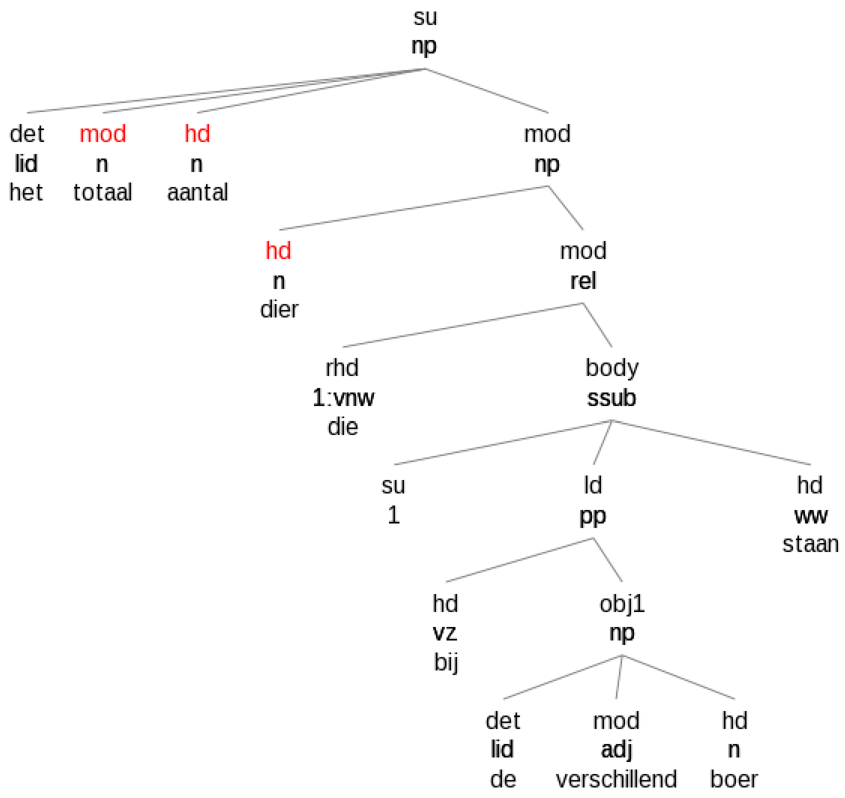


Figure 9: Three nouns – DEF-NS-Ns-NP-PL

- *miljoen* (million)
- *rij* (row)

(21) In Arnhem worden de enorme hoeveelheid antwoordbladen
in Arnhem become the enormous amount reply cards
gesorteerd
sorted

In Arnhem the huge amount of reply cards is sorted [WS-U-E-A-0000000043.p.36.s.4]

(22) Op de ruïnes van Hiroshima en Nagasaki bouwden de 127 miljoen
at the ruines of Hiroshima and Nagasaki built the 127 million

onderdanen een economische wereldmacht
subjects an economic world power

The 127 million subjects have built an economic world power on the ruins of Hiroshima and Nagasaki [WR-P-P-I-0000000254.p.5.s.1]

(23) De onderste rij panelen worden horizontaal verbonden door een
the bottom row panels be horizontally linked by a
hoge horizonlijn
high skyline

The lowest row of panels are connected by a high skyline [wiki-9720.p.36.s.2]

Note that the claim wrt type *tiental, aantal* in (Haeseryn et al., 1997) is said to hold for nouns expressing approximations, rough amount (benaderende hoeveelheid, onbepaalde hoeveelheid. In some sense this is the case as well for *127 million*, which will not be 127,000,000, but something inbetween 126,500,000 and 127,499,999. And what with *row*? The example provided (Haeseryn et al., 1997, p.1149) is with an indefinite article, and the hit mentioned above is with a definite one. But Google provides examples with indefinite articles as well, as in (24).

(24) a. Het maakt niet uit of er een rij mensen staan.PL te
it makes not out whether there a row people stand to
wachten
wait

It doesn't matter wheter a row of people is waiting (Google.be)

b. Een rij mensen die de polonaise doen.PL
a row people who the conga do

A row of people doing the conga (Google.be)

Assignment 1: "een van"-constructions (Haeseryn et al., 1997, pp.1142-1143):

(25) Een van hen die het kan/kunnen weten, is Dirk.

(26) Hij is een van de taalkundigen die daar wel eens over geschreven heeft/hebben.

(27) Een van de onderzoekers die bijdroeg/bijdroegen aan de ontwikkeling van het geneesmiddel, heeft een prijs gekregen.

(28) Een van de laatsten die het schip verlieten/?verliet was de purser.

(29) Dat was een van de eerste dingen die mij opvielen/?opviel.

Try to find such cases in LASSY, CGN, and/or SoNaR. (Search for 'een' both as number and as article) What about cases like (28) and (29)?

Assignment 2: Have a look at all the nouns that may appear as first noun in the binominal constructions discussed in thsi section (Table 5) Where do they belong in Table 1? And why?

Assignment 3: As mentioned above, there is another type of complex NP that, according to Haeseryn et al. (1997), can come with either a singular of plurel finite verb: *dit/dat soort mensen is/zijn niet te vertrouwen* 'this/that type of people is/are not to be trusted'.
Try to find examples in the corpora with such a demonstrative pronoun. Can *soort* be replaced by another noun, still showing the same behaviour?

5 GrETEL case study II: A treebank-based investigation of IPP verbs

5.1 Infinitivus Pro Participio

If a main verb occurs in the perfect tense, it appears as a past participle, cf. *gezeten* 'sat' in (30a).
By contrast, if a verb in the perfect tense selects another infinitive, i.e. in its serial use, it sometimes occurs as an infinitive instead of the (expected) past participle, such as *zitten* 'sit' in (30b) (Haeseryn et al., 1997, p.954). This

phenomenon is known as *Infinitivus Pro Participio* (IPP) or *Ersatzinfinitiv* (lit: 'substitute infinitive').

(30) a. Hij heeft urenlang op die bank *gezeten*.
he has hours-long on that bench sat
'He has been sitting on that bench for hours.'

b. Hij heeft *zitten* (te) slapen.
he has sit (to) sleep

c. * Hij heeft *gezeten* te slapen.
he has sat to sleep
'He has been sleeping.'

It is possible to differentiate between constructions that obligatorily show the IPP effect (30), constructions that optionally trigger IPP (31), and constructions in which IPP is not possible (32).

(31) a. De politie heeft de snelheidsmaniak *proberen* in te halen.
the police has the speed-maniac try in to overtake

b. De politie heeft *geprobeerd* de snelheidsmaniak in te halen.
the police has tried the speed-maniac in to overtake
'The police has tried to overtake the speed merchant.'

(32) a. * Hij heeft hem het raampje dicht *vragen* te doen.
he has him the window-DIM closed ask to do

b. Hij heeft hem *gevraagd* het raampje dicht te doen.
he has him asked to window-DIM close to do
'He asked him to close the little window.'

There are several lists of Dutch IPP verbs available in the literature, e.g. Rutten (1991), Haeseryn et al. (1997), and Klooster (2001). While there is consensus on the most common verbs, the authors disagree regarding the occurrence of IPP for some verbs and regarding the optionality of the phenomenon. Moreover, none of the authors claim to provide an exhausitive list, even though the set of IPP triggers is assumed to be a limited set of verbs.

Table 8 presents a list of IPP verbs based on **?**)pp.946-1082]ANS1997, Rutten (1991), and Klooster (2001).[27] The verbs indicated with a '+' obligatorily occur as IPP (according to the author(s) mentioned in the columns). Verbs indicated with a '–' cannot occur in IPP constructions, and verbs indicated with '±' optionally occur as IPP verbs. If a source does not mention the behaviour of a certain verb regarding IPP, it is indicated with '/'. The top part of the table lists the verbs which the authors agree upon regarding IPP. The middle part of the table contains the verbs on which two of the authors agree, while the bottom part of the table lists the verbs which were labelled differently by the three authors.

LEMMA	RUTTEN	HAESERYN ET AL.	KLOOSTER	TRANSLATION
(be)horen	+	+	+	ought to
blijven	+	+	+	stay, remain
dienen	+	+	+	be obliged to
doen	+	+	+	do, make
gaan	+	+	+	go, will
hoeven	+	+	+	need to
komen	+	+	+	come
kunnen	+	+	+	can, be able to
laten	+	+	+	let
liggen	+	+	+	lie
moeten	+	+	+	must, have to
mogen	+	+	+	may, be allowed to
staan	+	+	+	stand
vinden	+	+	+	find
weten	+	+	+	know (how to), remember
willen	+	+	+	want
zien	+	+	+	see

continued on next page

[27] In contrast to Rutten (1991) and Klooster (2001), Haeseryn et al. (1997) do not provide a list of IPP verbs. For some verbs, it is explicitely mentioned whether they occur in IPP constructions or not. For other verbs, the information in Table 8 is based on examples used in **?**)pp.946-1082]ANS1997.

Lemma	Rutten	Haeseryn et al.	Klooster	Translation
zien	+	+	+	manage
zitten	+	+	+	sit
zullen	+	+	+	will
beginnen	±	±	±	begin
helpen	±	±	±	help
leren	±	±	±	learn, teach
menen	±	±	±	mean, think
proberen	±	±	±	try
trachten	±	±	±	try
dreigen	±	–	–	threaten
durven	±	+	+	dare
hangen	/	+	/	hang
hopen	–	–	±	hope
horen	+	/	+	hear
lijken	+	/	/	seem
lopen	+	+	/	walk
pogen	/	±	±	try
schijnen	+	–	–	seem
voelen	+	/	+	feel
wagen	±	/	±	dare
weigeren	±	±	–	refuse
zijn	/	+	+	be in the activity of
begeren	–	/	±	desire
blijken	+	/	–	appear
plegen	+	–	±	be used to
vermogen	/	+	±	be in the power to, be able to
vrezen	–	/	±	fear
wensen	–	/	±	wish

Table 8: Dutch IPP verbs

The table shows that there are several verbs of which the IPP status is not agreed upon by the authors. A corpus-based study can help to overcome this

problem. The corpus study presented in the following sections aims, on the one hand, to check whether the verbs mentioned in the descriptive study indeed occur as IPP verbs in the data, and whether the data contain IPP verbs not mentioned in the literature.
On the other hand, the treebank data will provide a general idea of the frequency of IPP constructions, as well as a more detailed account of verbs which optionally appear as IPP verb.

5.2 Identifying IPP constructions with GrETEL

The IPP verbs in the treebanks were retrieved by means of GrETEL, as described in section 2. The automatically generated query in (2) was manually adapted to (5), repeated in (33), in order to include all IPP constructions in the treebanks.

(33)
```
//node[@cat and node[@rel="hd" and @pt="ww" and
(@lemma="hebben" or @lemma="zijn")] and node[@rel="vc"
and @cat="inf" and node[@rel="hd" and @pt="ww"]
and node[@rel="vc" and (@cat="inf" or @cat="ti" or
@pt="ww")]]]
```

Besides IPP constructions, the corresponding constructions with a past participle were extracted as well, in order to investigate the verbs that optionally appear as IPP . Those constructions, in which a verb of the perfect selects a past participle (PSP) followed by a (*te*)-infinitive, are henceforth referred to as *PSP constructions*. Extracting PSP constructions was done using the query in (34) which is very similar to the query for IPP constructions in (33): Only the phrasal tag of the VC node was changed from infinitival to participial (`@cat="ppart"`).

(34)
```
//node[node[@rel="hd" and @pt="ww" and (@lemma="hebben"
or @lemma="zijn")] and node[@rel="vc" and @cat="ppart"
and node[@rel="hd" and @pt="ww"] and node[@rel="vc" and
(@cat="inf" or @cat="ti" or @pt="ww")]]]
```

A graphical representation of the queries in (33) and (34) is presented in Figure 10.

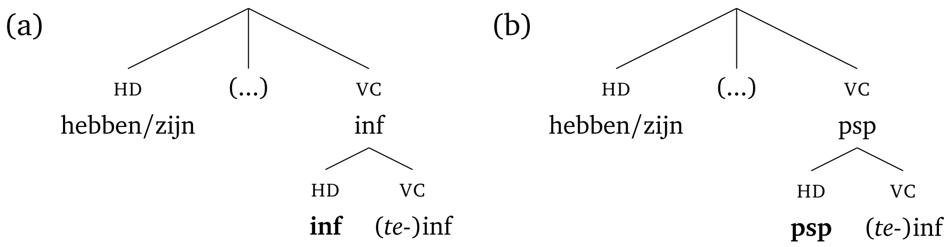


Figure 10: Query trees for IPP (a) and PSP (b) constructions

Note that in verb-initial clauses the finite verb and the other verbs are not necessarily adjacent. This is indicated in Figure 10 by (. . .). As XPath performs a *greedy* search, it is not necessary to specify this in the XPath expressions (see section 2).

Table 9 presents the general counts for both IPP constructions and PSP constructions, after manually filtering out erroneous results.[28]

	CGN		LASSY	
	# hits	% hits	# hits	% hits
IPP	792	96.2	309	71.2
PSP	31	3.8	125	28.8
TOTAL	**823**	**100**	**434**	**100**

Table 9: IPP and PSP constructions in CGN and LASSY Small

The IPP verbs are a subset of the verbs selecting a (*te*)-infinitival complement. Since there are 23.566 occurrences of such constructions in CGN and 16.448

[28] Sentences containing annotation errors or disfluencies were excluded from the data set, such as *ja nogal wiedes ja want je hebt een tijdje daar in die diepvrieskist daar zitten ... [CGN, fna000724__19],* which lacks an explicit second infinitive but has a tag for it.

in LASSY,[29] the IPP verbs only account for 3.4% of the verbs taking a (*te*-)-infinitival complement in CGN, and for 1.9% in LASSY. Still, IPP turns out to be a common phenomenon in a small set of frequently occurring verbs. Furthermore, the results in Table 9 reveal that in both treebanks IPP constructions occur more often than PSP constructions.

5.3 Typology of IPP verbs

After extracting the relevant constructions from the corpus, the IPP verbs were divided into *subject raising*, *object raising*, and *subject control* verbs, based on the criteria formulated in Sag et al. (2003), indicating which verb is able to select complements of its own, as well as its ability to assign semantic roles. No object control verbs (e.g. *overtuigen* 'persuade') have been detected as IPP. The resulting typology is presented in Table 10.[30] For each lemma, it is indicated how many times it appears as the selector of a (*te*)-infinitival complement (INF SEL), the frequency of its occurrence as IPP, the type of its VC complement if it appears as IPP, and its translation. The treebank counts in the table are the sums of the results obtained from the CGN and LASSY treebanks. In order to compare the treebank results with the descriptive study from section 5.1, the IPP verbs listed in the literature but not included in the treebank results were included in the table as well if their occurrence as IPP verb is confirmed in data from the Internet (using Google Search).[31] Those verbs have frequency 0 in the IPP column in Table 10. Optional IPP verbs are indicated with an asterisk (*).

[29] The constructions in which a verb selects a (*te*)-infinitival complement can be found by means of the following query: `//node[node[@rel="hd" and @pt="ww"] and node[@rel="vc" and (@cat="ti" or @cat="inf" or (@pt="ww" and @wvorm="inf"))]]`.

[30] Only the IPP verbs are listed, so verbs occurring in PSP constructions but not as IPP are not presented.

[31] `www.google.be/www.google.nl`

Lemma	inf sel	ipp	vc type	Translation
Subject raising verbs				
kunnen	8691	213	inf	can
moeten	7268	191	inf	have to
gaan	3748	188	inf	go, will
blijven	553	29	inf	continue
mogen	1393	29	inf	may
beginnen *	306	12	(te) inf	start, begin
hoeven	326	7	(te) inf	have to
zullen	6717	1	inf	will
dienen	176	1	te inf	have to
plegen	4	0	te inf	tend, be in the habit of
(be)horen	14	0	te inf	ought to
blijken	227	0	te inf	turn out
lijken	176	0	te inf	seem
schijnen	50	0	te inf	appear
dreigen	61	0	te inf	be on the point of
Object raising verbs				
laten	1428	154	inf	let
zien	363	36	inf	see
horen	74	19	inf	hear
doen	180	6	inf	do
helpen *	50	1	(te) inf	help
leren *	8	1	(te) inf	teach
voelen	15	0	inf	feel
Subject control verbs				
willen	2731	53	inf	want
zitten	379	44	(te) inf	sit
komen	383	27	(te) inf	come
leren *	91	19	(te) inf	learn
weten	156	19	(te) inf	know (how to)
staan	150	15	(te) inf	stand

continued on next page

Lemma	inf sel	ipp	vc type	Translation
zijn (wezen)	1011	13	inf	be in the activity of
proberen *	496	9	(te) inf	try
durven	90	8	(te) inf	dare
lopen	28	5	(te) inf	walk
trachten *	41	1	te inf	try
hangen	0	0	(te) inf	hang
liggen	29	0	(te) inf	lie
pogen *	8	0	te inf	try
weigeren *	71	0	te inf	refuse
begeren *	0	0	te inf	desire, want
hopen *	65	0	te inf	hope
menen *	19	0	te inf	mean, intend
vermogen *	0	0	te inf	be able to
wagen *	2	0	te inf	risk
wensen *	42	0	te inf	wish
SUM	**37662**	**1101**		

Table 10: Treebank-based typology of IPP verbs

The verb *leren* occurs twice in Table 10, since it has different meanings. If it has the meaning 'teach', it is a raising verb (35a), but it is a control verb if it denotes the meaning 'learn' (35b).

(35) a. 'k heb mijn kleine kinderen daar ook **leren** zwemmen .
I have my small children there also teach swim
'I have also taught my little children how to swim over there' [CGN, fva400659_44]

b. In 2001 heb ik saxofoon **leren** spelen .
in 2001 have I saxophone learn play
'In 2001 I learned to play the saxophone.' [LASSY, dpc-qty-000936-nl-sen.p.36.s.2]

The results of the treebank search furthermore reveal that the raising verbs are typical IPP verbs, especially the subject raisers *gaan* 'go', *moeten* 'have to', and *kunnen* 'can', cf. (36).

(36) Pas nu hebben we dat ook **kunnen** zien in de hersenen.
Only now have we that also can see in the brains
'Only now we have been able to see that in the brains.' [LASSY, dpc-ind-001634-nl-sen.p.16.s.5]

Based on the literature, most of the raising verbs obligatorily appear as IPP, i.e. all modal and evidential subject raisers, and the perceptive and causative object raising verbs. Due to data sparseness, not all of those verbs occur as IPP triggers in the treebanks, such as the evidentials *blijken* 'turn out', *lijken* 'seem', and *schijnen* 'appear', the modal *(be)horen* 'ought to', and the aspectual *plegen* 'be in the habit of' and *dreigen* 'be on the point of'.[32] A Google search reveals that they indeed do occur in IPP constructions, cf. (37).[33]

(37) a. Indien binnen de 14 dagen het matras niet de juiste keuze heeft **blijken** te zijn, dan moet er gekeken worden naar de problemen

b. . . . een klein stukje authentiek Brabant, waar de tijd stil heeft
. . . a small piece.DIM of Brabant, where the time still has
lijken te staan.
seem.INF to stand
'. . . a small piece of Brabant, where time seemed to stand still.'

c. mogelijk omdat licht in die rustige 20e eeuw minder belangrijk heeft **schijnen** te zijn.

d. Belanghebbende heeft echter geen feiten aangevoerd die tot het oordeel zouden moeten leiden dat de bank heeft begrepen of heeft **behoren** te begrijpen dat ...

[32] Note that only subject raising *dreigen* 'be on the point of' is an IPP verb. Its subject control homonym, meaning '(consciously) threaten', always appears as a past participle in the perfect tense, e.g. *Zowel Rwanda als Burundi hebben gedreigd hun buurland Congo binnen te vallen.* 'Both Rwanda and Burundi have threatened to invade their neighbouring country Congo.' [LASSY, WS-U-E-A-0000000230.p.11.s.3].

[33] All examples are found using Google.nl [17-07-2014].

e. De oude strever naar macht droeg zijn onttroning met de gelatenheid waarmee hij alle tegenslagen had **plegen** te ontvangen.

f. Hij vertelt het zonder emotie, alsof hij dit al lang heeft **voelen** aankomen.

g. Die ene kus had geleid tot een volgende en was ontvlamd in een vuurzee die hen allebei had **dreigen** te overspoelen.

Raising verbs that optionally trigger IPP are the benefactive object raisers *leren* 'teach' and *helpen* 'help', and the aspectual subject raiser *beginnen*.
The set of subject control verbs is more heterogeneous with regard to the IPP effect. It contains verbs that obligatorily appear as IPP, such as aspectual motion and position verbs such as *lopen* 'walk' and *staan* 'stand' (38). Furthermore, it contains verbs which occur in both IPP and PSP constructions, e.g. *proberen* 'try' (39a-39b), and verbs that do not allow IPP, such as *besluiten* 'decide'.

(38) Met wat meer geluk hadden we hier **staan** juichen.
With some more luck had we here stand cheer
'With a bit more luck we would have been cheering.' [LASSY, WR-P-P-H-0000000020.-p.14.s.8]

(39) a. ik heb m'n nicht **proberen** te bellen want die uh ...
I have my cousin try to call because that one uh
'I have tried to call my cousin because she erm ...' [CGN, fna000628__73]

b. daar hebben we toen **geprobeerd** te bellen ...
there have we then tried to call
'Then we have tried to call there' [CGN, fna000260__277]

Several verbs belonging to this category are mentioned in the literature as IPP verbs, but were not encountered in the treebanks. A Google search indicates that they are all optional IPP verbs (40),[34] except for *hangen* 'hang' and *liggen* 'lay', which always appear as IPP if they select an infinitival complement. However, for most verbs, their occurrence as an IPP verb is marginal. In the case of *begeren* 'desire' and *vermogen* 'be able to', IPP constructions were only encountered in archaic and very formal contexts.

[34] All examples except (40p) are found using Google.be [17-07-2014].

(40) a. . . . die Hij had begeren te redden uit de machtige klauw van Satan.

b. De rest van de wereld die hij had begeerd te zien.

c. Ze was dat boek gaan lezen omdat ze er iets uit had hopen te leren.

d. Ik had gehoopt te kunnen antwoorden op je vraag, maar helaas...

e. Van de gekwelde man die ze de dag daarvoor had menen te zien, was niets meer over.

f. Ik heb gemeend te kunnen vertrouwen op de overheid en het resultaat is dat ik met lege handen sta.

g. zit daar een diepere bedoeling achter, die ik niet heb vermogen te doorgronden?

h. . . . waardoor haar zieleblik nog nooit had vermocht te staren.

i. . . . het arme schaap faalde tijdens haar laatste examenrit omdat ze - hoe durft ze - een stopbord had weigeren te negeren . . .

j. de patiënt wenst niet deel te nemen en heeft geweigerd te tekenen

k. Omdat ik een ADSL-aanvraag bij Telfort heb wagen te annuleren . . .

l. Hij is een van de eerste wetenschappers die het heeft gewaagd te verdedigen dat dieren emoties hebben, mogelijk net als mensen.

m. Dat is de reden waarom ik mijn patronaat heb wensen te geven en samen met u deze conferentie vanavond over België in het hart van de Verenigde Naties heb

n. De wetgever heeft gewenst te voorkomen dat een onderneming zou starten met een te beperkt startkapitaal . . .

o. onmiddellijk nadat de bewakingscentrale de contactpersoon heeft pogen te bereiken.

p. Mogelijk heeft Milosevic gepoogd de Britse premier Blair te laten vermoorden. [LASSY, WR-P-P-H-0000000044.p.7.s.1]

q. waar ik de afdeling allereerst heb opgebouwd en het klachtregistratiesysteem heb helpen te ontwikkelen.

r. Deze reactie is een oeroud mechanisme dat ons heeft geholpen te overleven tijdens acuut gevaar.

No IPP examples of *vinden* 'find, think' and *vrezen* 'fear' could be found. Therefore, those verbs were not included in Table 10. Klooster (2001) claims that *vrezen* is an optional IPP verb. *Vinden* is said to be an obligatory IPP verb by Rutten (1991), Haeseryn et al. (1997) and Klooster (2001). Only Haeseryn et al. (1997) provide some examples (41), but they also mention that such constructions hardly occur in the perfect tense (Haeseryn et al., 1997, p.1022).

(41) a. Ze hebben hem op de grond vinden liggen.
they have him on the ground find lie
'They have found him lying on the ground.'

b. Ik heb dat nooit bij u vinden passen.
I have that never with you find suit
'I have never thought it suited you.'

Finally note that there are no verbs detected as IPP verbs that were not mentioned in the literature on the phenomenon.

5.4 Conclusion

Starting from a division into subject raising, subject control, and object raising verbs (Sag et al., 2003), a distinction is made along syntactic lines to account for the differences and similarities between IPP verbs.

The classification is supplemented with quantitative information to provide a general idea of the frequency of IPP verbs on the one hand, as well as a more detailed account of verbs which optionally appear as IPP on the other hand.

The classification furthermore shows that subject and object raising verbs are typical IPP verbs, whereas subject control verbs can be subdivided into verbs that obligatorily occur as IPP, optionally occur as IPP, or cannot occur as IPP. No object control verbs appear as IPP.

The list of IPP verbs presents the verbs that (obligatorily or optionally) occur as clustering verbs if they select an infinitival complement. Moreover, it provides us the set of verbs selecting a *te* infinitive that are always clustering, i.e. the verbs in Table 10 without an asterisk and with VC TYPE *(te)-inf* or *te inf*. That set allows to extract the clustering constructions with a *te* infinitive that do not occur in the perfect tense.

Assignment 1 Look for constructions with a tense auxiliary followed by two verbs that never show the IPP effect. For example*(dat) hij is/heeft opgehouden te vechten* 'that he has stopped fighting'.

Assignment 2 What about the choice of the auxiliary in constructions like *(dat) hij is/heeft opgehouden te vechten*. Do you find the same verbs for both *hebben* 'have' and *zijn* 'be'? Why (not)? Is it a coincidence or something more structural?

6 Advanced use cases

6.1 Complex proper names

Suppose you want to know which strings in CGN are marked as complex proper names, e.g. *Jan Jansen, Amsterdam ArenA, Den Haag, 's Hertogenbosch*. How would you proceed to get such data?[35]
Do you expect to find constructions consisting of three parts as well, such as *Memorial van Damme* and *Egmond aan Zee* ? Does the advanced search option provide you with an option allowing you to filter out clearly wrong examples, such as expressions in English or German? And what kind of expressions are marked as being 'not in Dutch'?

6.2 Same word, other meaning?

Although VL-Dutch and NL-Dutch are *in se* the same language, some words just have a slightly different meaning, or come with another article.

[35] Tip: have a look at the way such complex names are tagged and parsed in step 2.

Try to find out for the noun *wagen* (a type of vehicle) whether it refers to the same type of vehicle in Belgium and the Netherlands.

6.3 Full PP

Suppose that you want to look in CGN for full PPs in sentences containing Direct Objects (DOs) and Indirect Objects (IOs) as well. The PPs you are looking for could express a location, a direction, a duration, ... They are not part of the DO or IO.

- How are DOs, IOs and PPs expressed in the treebank under consideration? How can you figure this out without consulting the (syntactic) manual?
- Find two ways to search for such PPs.
- How to generalize over verb-initial and verb-final clauses?
- Should the full treebank be taken into account? Why (not)?[36]

6.4 Separable verbs

Look for instances of separable verbs, including the cases with intervening words.

(42) *aanhouden* ‘continue’

a. ..., dat de storm aan zal houden

b. De storm houdt aan

(43) *iemand aanhouden* ‘to stop someone’

a. ..., dat hij haar aan moet houden

b. Hij houdt haar aan

[36] Tip: Have a look at the description of the subcomponents.

(44) *iemand houden aan X* ‘to make someone keep his/her X’
a. ..., dat hij haar aan haar belofte houdt
b. Hij houdt haar aan haar belofte

So you want instances of (42) and (43), while you are not interested in cases like (44).

- How to proceed?
- How many instances do you find in LASSY of the verb *aanhouden* (split forms)?
- Look for a non-split instance of any separable verb. Have a look at the tree structure of one of the examples (click in Step 6 on a sentence ID, the tree structure will open in a new window).
- Have a look at the tree structure of a split instance as well.
- Compare the characteristics of the separable verb in both instances (esp. `root`, `lemma`, `rel` and `pt`). You can see the value of the `root` and detailed POS tags by hovering over the nodes in the tree with the cursor.
- Do you have any idea how you could find all instances of the separable verb *aanhouden* in one search action, using the possibility to adapt the XPath-query (Step 5, Advanced mode)? Give it a go!

6.5 Types of indirect objects

In Dutch you have constructions like

(45) a. Jan gaf Marie een boek
b. Jan gaf aan Marie een boek
c. Jan gaf een boek aan/*voor Marie

(46) a. Jan schonk Marie een borrel in

b. Jan schonk voor Marie een borrel in

c. Jan schonk een borrel in voor/*aan Marie

The constructions under (45) contain a *meewerkend voorwerp* 'indirect object' (IO), those under 46 a *belanghebbend voorwerp* also 'indirect object'. But the two types together are also called *meewerkend voorwerp* 'indirect object'.[37] Note that it is quite often not that easy to distinguish a *meewerkend voorwerp* (narrow meaning) and a *belanghebbend voorwerp* from other NPs and PPs. In LASSY NPs and PPs that function as complements (i.e. NPs and PPs that are selected by a verb) and combine with either *aan* or *voor* are considered indirect objects (OBJ2).

- Count the bare IOs in LASSY (i.e. those without prepositions *aan* or *voor*).
- Count those with *aan* or *voor* as well.
- How many instances did you find?
- Could you extract them all in one go? Why (not)?
- Since the number or results is not that high, have a look at the structures found for OBJ2s realized as PPs. Are there clear mistakes? Be aware that a realisation without preposition does not need to be a rather obvious one, what counts is whether it is possible or not.
- What does this tell you?

Wait ... did you really find all IOs?

(47) ik geef een boek aan het kind

(48) ik geef een boek aan de kinderen

(49) ik geef een boek aan de jongen

[37] *Meewerkend voorwerp* thus has two meanings; *indirect object* even three.

(50) ik geef een boek aan de kleine kinderen

(51) ik geef een boek aan Marie

(52) ik geef een boek aan juf Marie

(53) ik geef een boek aan Jan Janssen

(54) ik geef een boek aan jou

(55) ik geef een boek aan jouw zusje

The same holds for the subject, the direct object, the choice of verbs (*geef* vs *heb gegeven*), etc. Try to find the optimal generalization, both in the basic and the advanced search mode.

6.6 *Hen, hun, zullie, hullie* (hebben het gedaan)

According to Haeseryn et al. (1997), *hen* is preferred over *hun* when used as object from a stylistic point of view (Haeseryn et al., 1997, pp.247-248). *Hullie* and *zullie* are not even mentioned.

(56) ik geef hen/hun een boek

(57) ik zie hen/hun

(58) Als wij beter voetballen dan hullie, halen we meer punten dan zullie (toegeschreven aan Johan Cruyff)

See also `http://www.taalcanon.nl/vragen/is-hun-hebben-zeggen-echt-zo-dom/` and `http://www.collegenet.nl/collegedoc/index.php?document=4693`. Have a look at the occurrence of *hen* en *hun* in both CGN and LASSY. Does this confirm what is said in the literature? What about the distribution over written and spoken language? Tip: you may want to take into account the nature of the various subcomponents. And what about Flanders versus the Netherlands?

6.7 Gender of nouns

In the corpora included in GrETEL all nouns are said to be neuter or non-neuter, as most people in the Netherlands, contrary to Flanders, do not know which non-neuter nouns are masculine of feminine. But even then, there are some mismatches, words that are considered neuter in Flanders, but non-neuter in the Netherlands, or the other way around while the words have the same meaning (*pad* is therefore not a correct example, *de pad* (non-neuter) being a toad, and *het pad* (neuter) a path. But a word like *filter* 'filter, colander' would do.)
How is this encoded in the copora, i.e. what specific feature do these words show?

Assignment Try to find 5 words in CGN and in LASSY that come with this specific feature.

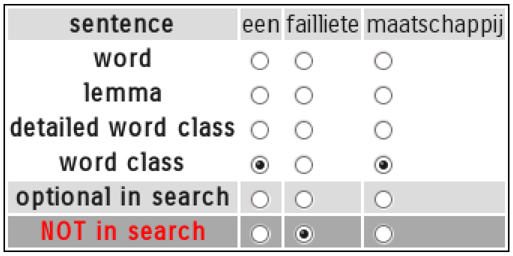
Tip 1: You should look for constructions with *een*, in other contexts the gender-issue is solved by the choice of the definite article, the pronoun, ...

Tip 2: Try and find a noun (on the web) that behaves differently as far as gender is concerned in Flanders and the Netherlands. Look for the word in CGN and/or LASSY. In the results (step 6) click on a sentence ID to view a tree structure. Hover over the word under consideration with the cursor to inspect the detailed POS tag. What does it show instead of non-neuter (`zijd` for *zijdig*) or neuter (`onz`, for *onzijdig*).

The Alpino parser was developed in the Netherlands. Is that clear from the analysis of constructions of this type in step 5? How can this be avoided (advanced mode)?

6.8 The NOT option: A restricted set of NPs

Try to find NPs without adjectives modifying the noun, i.e. not *een failliete maatschappij* ‘a bankrupt society’, but only *een maatschappij* ‘a society’. By using the advanced search mode you can use the `not in search` option in the matrix (step 3), see Figure 11.



sentence	een	failliete	maatschappij
word	○	○	○
lemma	○	○	○
detailed word class	○	○	○
word class	●	○	●
optional in search	○	○	○
NOT in search	○	●	○

Figure 11: een <empty> maatschappij

For the DPC-subcomponent (LASSY) this returns over 16,000 hits. Have a look at them.
Some questions:

- dpc-bal-001238-nl-sen.p.54.s.4: This sentence contains *de aanpak van dit mondiale vraagstuk* (the way to deal with this global problem). Why is this correct?

- dpc-bmm-001078-nl-sen.p.4.s.3: This sentence does not seem to contain any ‘bare NPs’. So why is it returned as a hit?

6.9 Only definite articles

Search for NPs containing the definite articles *de* or *het* ‘the’, while excluding the indefinite article *een* ‘a’. What is an easy way do do this, once you are somewhat familiar with the advanced search mode, i.e. with the possibility to edit XPath expressions?
Have a look at the XPath expression in (33). It contains an ‘or’-operator.

Whereas the original, automatically generated XPath expression would have contained either `@lemma="hebben"` or `@lemma="zijn"`.

Can a similar generalisation be used to find both *de* and *het* in one search?

7 Reading materials

7.1 About GrETEL

Augustinus, L., V. Vandeghinste & F. Van Eynde. 2012. Example-Based Treebank Querying. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, 3161–3167. Istanbul: European Language Resources Association (ELRA)

Augustinus, L., V. Vandeghinste, I. Schuurman & F. 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)*, 423–428. Oslo: NEALT Proceedings Series 16

More specialised:
Vandeghinste, Vincent & Liesbeth Augustinus. 2014. Making Large Treebanks Searchable. The SONAR case. In Marc Kupietz et al. (eds.), *Proceedings of the LREC2014 2nd workshop on Challenges in the management of large corpora (CMLC-2)*, 15–20. Reykjavik

7.2 Using GrETEL

Augustinus, L. & F. Van Eynde. 2012. A Treebank-based Investigation of IPP-triggering Verbs in Dutch. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*, 7–12. Lisbon: Edições Colibri

Van Eynde, F., L. Augustinus, I. Schuurman & V. Vandeghinste. 2014. Het verrassende resultaat van een copulativiteitspeiling. In F. Van de Velde, H. Smessaert, F. Van Eynde & S. Verbrugge (eds.), *Patroon en argument. een dubbelfeestbundel bij het emeritaat van william van belle en joop van der horst*, 47–62. Leuven: Universitaire Pers

Augustinus, L. & F. Van Eynde. 2015. Looking for Cluster Creepers in Dutch Treebanks. *Dat we ons daar nog kunnen mee bezig houden. Computational Linguistics in the Netherlands Journal* 4

7.3 The corpora

CGN:
Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J.-P. Martens, M. Moortgat & H. Baayen. 2002. Experiences from the Spoken Dutch Corpus Project. In Manuel Gonzalez Rodriguez & Carmen Paz Saurez Araujo (eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 340–347. Las Palmas

Hoekstra, H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman & T. van der Wouden. 2003. *CGN Syntactische Annotatie*. 77p

van der Wouden, T., I. Schuurman, M. Schouppe & H. Hoekstra. 2003. Harvesting Dutch trees: Syntactic properties of Spoken Dutch. In Tanja Gaustad et al. (eds.), *Computational Linguistics in the Netherlands 2002*, 129–141. Amsterdam: Rodopi

Schuurman, Ineke, Wim Goedertier, Heleen Hoekstra, Nelleke Oostdijk, Richard Piepenbrock & Machteld Schouppe. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ... In *Proceedings of the Fourth*

International Conference on Language Resources and Evaluation (LREC-2004), vol. 1, 57–60. Lisbon

LASSY:
van Noord, Gertjan, Ineke Schuurman & Vincent Vandeghinste. 2006. Syntactic Annotation of Large Corpora in STEVIN. In *Proceedings of the fifth international conference on language resources and evaluation (lrec-2006)*, vol. 1, 1811–1814. Genoa

van Noord, G., I. Schuurman & G. Bouma. 2011. *Lassy Syntactische Annotatie, Revision 19455*. Www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf

van Noord, G., G. Bouma, F. Van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang & V. Vandeghinste. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In P. Spyns & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: resources, tools and applications*, Springer

SoNaR:
Oostdijk, Nelleke, Martin Reynaert, Paola Monachesi, Gertjan van Noord, Roeland Ordelman, Ineke Schuurman & Vincent Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008)*, 1437–1444. Marrakech

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: resources, tools and applications*, 219–247. Springer

References

Abeillé, A. (ed.). 2003. *Treebanks. Building and Using Parsed Corpora*. Dordrecht: Kluwer.

Augustinus, L. & F. Van Eynde. 2015. Looking for Cluster Creepers in Dutch Treebanks. *Dat we ons daar nog kunnen mee bezig houden. Computational Linguistics in the Netherlands Journal* 4.

Augustinus, L. & F. Van Eynde. 2012. A Treebank-based Investigation of IPP-triggering Verbs in Dutch. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*, 7–12. Lisbon: Edições Colibri.

Augustinus, L., V. Vandeghinste, I. Schuurman & F. 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)*, 423–428. Oslo: NEALT Proceedings Series 16.

Augustinus, L., V. Vandeghinste & F. Van Eynde. 2012. Example-Based Treebank Querying. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, 3161–3167. Istanbul: European Language Resources Association (ELRA).

van der Beek, L. 2005. *Topics in Corpus-Based Syntax*: Groningen Dissertations in Linguistics dissertation.

van der Beek, L., G. Bouma, R. Malouf & G. van Noord. 2002. The Alpino Dependency Treebank. In *Computational Linguistics in the Netherlands 2001*, Rodopi.

Bouma, G. & G. Kloosterman. 2002. Querying dependency treebanks in XML. In *Proceedings of the Third international Conference on Language Resources and Evaluation (LREC-2002)*, Gran Canaria.

Bouma, G. & G. Kloosterman. 2007. Mining Syntactically Annotated Corpora using XQuery. In *Proceedings of the Linguistic Annotation Workshop (ACL 07)*, .

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M. van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Martinus Nijhoff/Wolters Plantyn 2nd edn.

Hoekstra, H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman & T. van der Wouden. 2003. *CGN Syntactische Annotatie*. 77p.

Huddleston, R. & G.K. Pullum. 2002. *The Cambridge Grammar Of The English Language*. Cambridge University Press.

Klooster, W. 2001. *Grammatica van het hedendaags Nederlands. Een volledig overzicht*. Den Haag: Sdu Uitgevers.

Lezius, W. 2002. TIGERSearch - Ein Suchwerkzeug für Baumbanken. In Stephan Busemann (ed.), *Proceedings of KONVENS-02*, Saarbrücken.

Meurers, D. & S. Müller. 2009. Corpora and Syntax. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, chap. 42, 920–933. Berlin: Mouton de Gruyter.

Nivre, J. 2008. Treebanks. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 1, chap. 13, 225–241. Berlin: Mouton de Gruyter.

van Noord, G. 2006. At Last Parsing Is Now Operational. In P. Mertens, C. Fairon, A. Dister & P. Watrin (eds.), *TALN 2006. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42.

van Noord, G., G. Bouma, F. Van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang & V. Vandeghinste. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In P. Spyns & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: resources, tools and applications*, Springer.

van Noord, G., I. Schuurman & G. Bouma. 2011. *Lassy Syntactische Annotatie, Revision 19455*. Www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf.

van Noord, Gertjan, Ineke Schuurman & Vincent Vandeghinste. 2006. Syntactic Annotation of Large Corpora in STEVIN. In *Proceedings of the fifth international conference on language resources and evaluation (lrec-2006)*, vol. 1, 1811–1814. Genoa.

Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J.-P. Martens, M. Moortgat & H. Baayen. 2002. Experiences from the Spoken Dutch Corpus Project. In Manuel Gonzalez Rodriguez & Carmen Paz Saurez Araujo (eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 340–347. Las Palmas.

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: resources, tools and applications*, 219–247. Springer.

Oostdijk, Nelleke, Martin Reynaert, Paola Monachesi, Gertjan van Noord, Roeland Ordelman, Ineke Schuurman & Vincent Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008)*, 1437–1444. Marrakech.

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

Rutten, J. 1991. *Infinitival Complements and Auxiliaries.*: University of Amsterdam dissertation.

Sag, I., T. Wasow & E. Bender. 2003. *Syntactic Theory. A Formal Introduction.* Stanford: CSLI Publications 2nd edn.

Schuurman, Ineke, Wim Goedertier, Heleen Hoekstra, Nelleke Oostdijk, Richard Piepenbrock & Machteld Schouppe. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ... In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, vol. 1, 57–60. Lisbon.

Van Eynde, F. 2009. A Treebank-driven Investigation of Predicative Complements in Dutch. In B. Plank, E. Tjong Kim Sang & T. Van de Cruys (eds.), *Computational Linguistics in the Netherlands 2009* LOT Occasional Series 14, 131–145. Utrecht.

Van Eynde, F., L. Augustinus, I. Schuurman & V. Vandeghinste. 2014. Het verrassende resultaat van een copulativiteitspeiling. In F. Van de Velde, H. Smessaert, F. Van Eynde & S. Verbrugge (eds.), *Patroon en argument. een dubbelfeestbundel bij het emeritaat van william van belle en joop van der horst*, 47–62. Leuven: Universitaire Pers.

Vandeghinste, Vincent & Liesbeth Augustinus. 2014. Making Large Treebanks Searchable. The SONAR case. In Marc Kupietz et al. (eds.), *Proceedings of the LREC2014 2nd workshop on Challenges in the management of large corpora (CMLC-2)*, 15–20. Reykjavik.

van der Wouden, T., I. Schuurman, M. Schouppe & H. Hoekstra. 2003. Harvesting Dutch trees: Syntactic properties of Spoken Dutch. In Tanja Gaustad et al. (eds.), *Computational Linguistics in the Netherlands 2002*, 129–141. Amsterdam: Rodopi.