SPOD

van Noord, Gertjan; Hoeksema, Jack ; Kleiweg, Peter; Bouma, Gosse

Published in:
Computational Linguistics in the Netherlands Journal

Computational Linguistics in the Netherlands Journal 10 (2020) 129-145 Submitted 10/2020; Published 12/2020

SPOD: Syntactic Profiler of Dutch

Gertjan van Noord* G.J.M.VAN.NOORD@RUG.NL
Jack Hoeksema* J.HOEKSEMA@RUG.NL
Peter Kleiweg* P.C.J.KLEIWEG@RUG.NL
Gosse Bouma* G.BOUMA@RUG.NL

* *University of Groningen, Groningen, Netherlands*

Abstract

SPOD is a tool for Dutch syntax in which a given corpus is analysed according to a large number of predefined syntactic characteristics. SPOD is an extension of the PaQu ("Parse and Query") tool (Odijk et al. 2017). SPOD is available for a number of standard Dutch corpora and treebanks. In addition, you can upload your own texts which will then be syntactically analysed.

SPOD will run a potentially large number of syntactic queries in order to show a variety of corpus properties, such as the number of main and subordinate clauses, types of main and subordinate clauses, and their frequencies, average length of clauses (per clause type: e.g. relative clauses, indirect questions, finite complement clauses, infinitival clauses, finite adverbial clauses, etc.). Other syntactic constructions include comparatives, correlatives, various types of verb clusters, separable verb prefixes, depth of embedding etc.

SPOD allows linguists to obtain a quick overview of the syntactic properties of texts, for instance with the goal to find interesting differences between text types, or between authors with different backgrounds or different age. In the paper, we describe the SPOD tool in some more detail, and we provide a case study, illustrating the type of investigations which are enabled and facilitated by SPOD.

Most of the syntactic properties are implemented in SPOD by means of relatively complicated XPath 2.0 queries, and as such SPOD also provides examples of relevant syntactic queries, which may otherwise be relatively hard to define for non-technical linguists.

SPOD is available via `https://www.let.rug.nl/alfa/paqu/spod`

1. Introduction

Statistical information is relevant for many domains of linguistic inquiry that use corpus methods, including L1 and L2 acquisition research, historical linguistics, psycholinguistics, sociolinguistics and text classification. A lot has been done with small, manually annotated corpora such as the Penn Treebank (Marcus et al. 1993), or Lassy Small (van Noord et al. 2013, van Noord and Bouma 2009, van Noord 2009), but with the availability of automatic parsers such as Alpino (Bouma et al. 2001, van Noord 2006), it is possible to obtain good quality syntactic information from automatically parsed corpora. Alpino output can already be searched using CLARIN-tools such as PaQu (Odijk 2015, Odijk et al. 2017, van der Wouden et al. 2015, Bouma 2017) and GrETEL (Augustinus et al. 2012). Moreover, unparsed corpora can be uploaded to PaQu, get parsed by Alpino, and then be queried. PaQu queries are not overly easy to use, but PaQu comes with a set of example queries, which help a lot to elucidate the query system, and can be modified to make new queries. These queries allow one to find examples of targeted structures in a corpus, but also provide statistical information regarding those structures.

SPOD is an extension available in PaQu, in which a large set of syntactic queries is available to characterize the syntactic make-up of a parsed corpus. For example, SPOD can be used if we like to know what percentage of clauses is interrogative, and, within the set of interrogative clauses, how many are WH-questions, how many yes/no questions, how many are indirect questions and how many are direct questions. For each subclass, SPOD provides information such as average length in

number of words. Likewise, we can obtain what the percentage of relative clauses is, and the relative size of the sets of free relative clauses and headed relative clauses. Again, for each subtype, we can inspect average length.

Similar information is provided for other constituents than clauses. What is the number of prepositional phrases (per million words) in the corpus, what is their average size, what part is adnominal modifier, what part is adverbial, what part is a prepositional complement and what part is a predicate? On the basis of such queries (and more), it is possible to analyse various corpora. In SPOD, a large set of such queries is available. The result will be a useful base line for more fine-grained syntactic studies of corpora.

An important aspect of SPOD, inherited from PaQu, is the option to upload your own text in plain text format (and some other formats). The text is tokenized and parsed automatically by Alpino and then is available in SPOD just like any of the other treebanks.

One application of SPOD is the analysis of texts by elementary and high school students, to gain a global insight in syntactic development across the school age. We use two corpora: a corpus of high school essays, collected by K. de Glopper, Groningen, and BasiScript, a corpus of texts by elementary school children, collected by Agnes Tellings, Radboud University.

There is already a large body of work in corpus linguistics targeting the first 5 or 6 years of L1 acquisition, but relatively little on the period between 6 and 18, the school age. By taking an inventory check of texts produced at various ages, a better insight in the process of syntactic development can be obtained. This is useful for our understanding not just of syntactic development, but also of usage preferences, reading and writing proficiency, and (especially for children of high school age) L2 acquisition. Regarding writing proficiency, we note that the corpus-De Glopper has metadata consisting of teacher grades for the essays. We can search for differences between highly graded and poorly graded texts.

Other applications of SPOD are not hard to come up with. For most applications, one feature of PaQu is very useful for our profiler, namely the option to use metadata from the corpus. This allows us to distinguish text types, or user categories, e.g. in the CHILDES corpus (already available in a parsed form) it is possible to distinguish the input of children and that of adults. In the high school data collected by De Glopper et al., we can distinguish age groups and school types, as well as males and females.

Data provided by SPOD are not only useful in comparing corpora syntactically, but may also be useful in other types of research. Fields where a syntactic profiler will be useful are text-type classification, stylometry, readability analysis, sociolinguistics and psycholinguistics. These fields currently rely heavily on the study of word distributions, but more syntactically oriented work is important as well (Biber 1993, Pander Maat et al. 2014, Jautze et al. 2013, Roland et al. 2007). SPOD facilitates the exploitation of syntactic properties in these research fields.

The availability of a large set of predefined syntactic queries in SPOD can simplify and automatize the study of parsed corpora for users who do not want to master the details of a query language. SPOD improves the user friendliness of PaQu by adding an interface allowing users to select options from a menu of predefined queries. In this way, PaQu-based studies such as Bouma (2017) and Odijk (2015) will become easier to carry out for regular working linguists.

The current paper provides an overview of the functionality, and the implementation of SPOD. In addition, we present a case study on the distribution of noun phrases, prepositional phrases, adjectival phrases and adverbial phrases in a variety of corpora. The goal of the case study is to illustrate how SPOD enables and facilitates the corpus-based study of syntactic phenomena.

2. Manually and automatically parsed corpora

Naturally, treebank search engines such as PaQu and SPOD are only as good as the annotations provided in the treebanks. The syntactic annotations in PaQu and SPOD are based on the guidelines originally put in place for the CGN corpus (van Eynde et al. 2000, Moortgat et al. 2000). These

guidelines were then slightly adapted and extended in the D-Coi and Lassy projects (van Eynde 2005, van Noord et al. 2019). The guidelines were meant to be theory neutral but it does mean that linguists working in a particular linguistic tradition must be aware of the annotation decisions used in the treebanks.

The manually verified and corrected treebanks sometimes still contain mistakes or inconsistencies. In a study about the annotation of the Alpino Treebank (van der Beek et al. 2002a), an overlap of dependency annotation of 93.1% percent was found between two annotators. After correction of clear mistakes, the annotation agreement increased to 94.6%. Most manually annotated materials have been checked by at least two annotators and in addition, various tools to find inconsistencies have been applied. Therefore, we expect that this number can be taken to be the lower bound of the quality of the dependency annotations in the manually corrected treebanks.

Obviously, the quality of automatically parsed corpora is lower than the manually corrected treebanks. The quality of the automatic annotations will vary with the nature of the texts. The accuracy of the dependency annotations for the LassySmall corpus ranges from 80% (legal texts) to 94% (books), with an avarage of almost 90% percent. For the Alpino Treebank (the "dbl" (newspaper) part of the Eindhoven corpus), the accuracy of the parser is in that range too (90.5%). Spoken language is much harder for the parser. For the manually annotated part of CGN, the parser obtains an accuracy over dependency annotations of 71.5%.

If the corpora of study are automatically parsed, results should be considered critically because of this reduced quality. In particular, this may be an issue in the case of texts that are further removed from standard Dutch, such as the texts produced by language learners. For this reason, it is important that SPOD not only provides the raw counts, but also provides direct access to the individual matched utterances so that researchers can assess the reliability of those counts. The access to the individual matches and the actual query is straightforward because of the functionality provided by PaQu. This link with PaQu might reduce some of the dangers noted in Odijk (2020).

Checking matches for correctness evaluates precision (are the examples found by the query indeed examples of the construction of interest?), but does not tell us anything about recall: perhaps good examples of the construction of interest were not parsed correctly, and therefore not found by the query. In order to judge recall, it may help to provide the parser with a number of sentences which illustrate the phenomenon of interest, e.g. via the on-line demo of Alpino at `urd2.let.rug.nl/~vannoord/bin/alpino`. Another approach, suggested in Bloem (2020) is to study the results for more general and less precise variants of the queries which return more hits, and to check for desired results in the more general queries which are not present in the final query.

In the overview below, we often present examples from a number of standard Dutch corpora, which are available in SPOD.

Lassy Small Lassy Small (van Noord et al. 2013) is a manually verified syntactically annotated corpus of written Dutch. It contains a variety of text types and is used below as a reference corpus of standard written Dutch.

CGN CGN is the "Corpus Gesproken Nederlands" (Corpus of Spoken Dutch) (Schuurman et al. 2003). Part of that corpus has manually verified syntactic annotations. If we refer to CGN below, we refer to this syntactically annotated part.

Wablieft Wablieft (Vandeghinste et al. 2019) is a corpus of simplified news articles from a Belgian newspaper in simple Dutch, for people who find ordinary newspapers too hard to read. The syntactic annotations are provided by the Alpino parser, and not manually verified.

Eindhoven The Eindhoven corpus (uit den Boogaart 1975) is a corpus of Dutch collected in the sixties. Even though one part of that corpus has manually verified syntactic annotations (known as the Alpino Treebank (van der Beek et al. 2002b)), below we refer to the full, automatically parsed version of this corpus.

Alpino Treebank The Alpino Treebank (van der Beek et al. 2002b) (van der Beek et al. 2002b) consists of manually verified syntactic annotations of the newspaper (cdbl) part of the Eindhoven corpus.

Although all corpora essentially follow the annotation guidelines described in van Eynde (2005) and van Noord et al. (2019), there are differences between the manually verified and automatically parsed corpora. The latter corpora contain additional information which is crucial for a small subset of the syntactic queries of SPOD. CGN predated the other corpora and in some cases the later Lassy guidelines differ in detail with the guidelines used for CGN. As a result, a few queries are regrettably not available for CGN.

3. Syntactic Queries

The core of SPOD is an inventory of more than hundred syntactic queries. For a given corpus, a user can select the relevant queries, or simply run the whole set. In this section, we shortly describe the various queries.

The result of running the profiler on a particular corpus starts with an overview of the size of the corpus in number of sentences and number of words, average number of words per sentence, average number of letters per word, and type/token-ratio. The notion 'sentence' here simply refers to the way in which the corpus is split in separate utterances, and need not correlate directly with the linguistic notion of root sentence, subordinate sentence or finite main clause.

In addition, the frequency overview for each of the individual syntactic queries is provided, with a hyperlink to all of the individual matches. This provides an intuitive and quick interface for the user to inspect the actual data in case of unexpected results.

The queries available in SPOD have been selected as follows. A linguist (second author) suggested the generic properties of corpora that linguists might be interested in. A computational linguist (first author) then came up with the appropriate query (sometimes after further consultation with the linguist). Over a period of two years, further queries have been added based on the experience of the linguist using the tool. Although we attempted to select a generically useful set of queries, the resulting list might appear to other linguists incomplete. The setup of the SPOD engine is such that adding further queries is rather straightforward, but at this point does require human intervention. We invite users of SPOD to come up with suggestions for further queries to be included in SPOD.

3.1 Words and word order

The SPOD profiler provides a distribution of part-of-speech labels, both for the twelve main POS-labels as well as for all detailed POS-labels.

A separate section provides detailed queries for verbs. Counts are provided for the number of fixed verbal expressions, the number of verb clusters, the number of passive verbs, impersonal passives, and the number of cross-serial verb cluster constructions.

The infamous Dutch cross-serial verb clusters (Huybregts 1984), illustrated in (1), appears to be syntactically complex. Indeed, in the Wablieft corpus of simplified Dutch the construction only occurs in 0.22% of the sentences. In the Eindhoven corpus, the construction occurs in 0.73% of the sentences. In Lassy Small, the proportion is 0.41%.

(1) [..] omdat ik Cecilia Henk de nijlpaarden zag helpen voeren
[..] because I Cecilia Henk the hippos saw help feed
"[..] because I saw Cecilia help Henk feed the hippos"

Authors are often advised not to use the passive construction if their text should be easy to read. This advice is taken to heart in Wablieft: the passive occurs in only 0.08% of the sentences whereas in Lassy Small the ratio is 2.59% (thus 32 times more frequent).

Furthermore, for verb clusters which contain a participle, SPOD provides frequencies of the so-called *green* order (verb cluster starts with participle as in example (2)) and *red* order (verb cluster ends with participle as in example (3)). Both examples are from Wablieft.The two orders are equally grammatical, but different regions show different preferences (Pauwels 1953, de Sutter et al. 2005).

(2) Tot nu kregen veel mensen de raad om thuis te blijven tot ze helemaal genezen
until now got many people the advice for home to stay till they completely healed
zijn
are
"Up to now many people were advised to stay at home until they were completely healed"

(3) Dat komt omdat veel Grieken de crisis hebben aanvaard
that comes because many Greeks the crisis have accepted
"That is because many Greeks have accepted the crisis"

A further section zooms in on verbs with a separable verb prefix. This part of SPOD is only available for automatically annotated corpora since the queries rely on additional information provided by the parser that is not available in manually annotated corpora. Information can be provided on the number of verbs with a separable verb prefix, and the proportion of cases in which the prefix is incorporated in the verb, or separated, both for finite and non-finite verbs.

If we compare the amount of verbs with separable verb prefixes in the Eindhoven corpus and in Wablieft, then we note that such verbs occur somewhat more often in the Eindhoven corpus (10974 out of 623092 words: 1.76%) than in Wablieft (31123 cases out of 2070574 words: 1.50%). In declarative, finite, main clauses, the finite verb cannot incorporate the separable verb prefix. And since the proportion of declarative main clauses is much higher for the Wablieft corpus, we should only compare the proportion of separated and non-separated verb prefixes for non-finite verbs. If we do this, we note that in Wablieft the proportion of incorporated verb prefixes (79%) is quite similar to the proportion in the Eindhoven corpus (75%).[1]

For automatically parsed corpora, information is provided on words that were unknown to the parser. Detailed counts are provided for words that the parser guessed were names, compounds or otherwise. For instance, the recent Wablieft corpus of news articles in simplified language contains over two million words. Of those, 66598 were unknown (2.89%). The parser decided that these consisted of 42193 names and 10393 compounds. In comparison, for the automatically parsed version of the Eindhoven corpus the proportion of unknown words is considerably larger, 3.92%.

3.2 Main and subordinate sentences

Information is provided on the distribution of main clause types in declarative sentences, WH-questions, yes-no-questions and imperatives. For subordinate sentences, a variety of counts is provided distinguishing finite and infinite subordinates sentences. For infinite subordinate sentences, further details are provided depending on the grammatical role of the clause. SPOD also provides the number of relative clauses and the number of free relatives.

As we might expect, the number of free relatives - a relatively complicated syntactic construction - is much less frequent in Wablieft (in 0.61% of the sentences) than in Lassy Small (1.63%) or the Eindhoven corpus (3.32%). Even in CGN, free relatives occur somewhat more often (1.18%). Here is an example from the Wablieft corpus:

(4) Wie zelf geen inkomen heeft om te overleven, kan bij het OCMW hulp vragen
who self no income has for to survive, can at the OCMW help ask

1. The latter percentages are not provided by SPOD.

"Who does not have an income to live on can ask for help at the OCMW"

A somewhat more syntactically involved section provides information on topicalization and extraction. For declarative sentences starting with a NP, SPOD provides information on the role of that NP. Is it a subject (the unmarked case) or does it have another grammatical role? In the Wablieft corpus, the unmarked case occurs 151190 times, whereas in only 10288 cases the NP has a different role. A few examples of the latter are given here:

(5) Dat weten ook de Verenigde Naties (Wablieft)
that know also the United Nations
"The United Nations know that as well"

(6) Dat gerecht maken Chinezen met eieren van een eend (Wablieft)
that dish make Chinese with eggs of a duck
"Chinese make that dish with duck eggs"

SPOD also provides the number of cases in which the first constituent of a declarative sentence is grammatically dependent on an embedded constituent. In Wablieft, only a single example is found which appears to be a mis-parse caused by a tokenization problem. In the Alpino Treebank and Lassy Small, no examples can be found. In spoken language, this construction does surface somewhat more often. The following examples are from CGN:

(7) en dat vrees ik dat ie wel gaat doen
and that fear I that he AFF goes do
"And that, I fear that he WILL do"

(8) ene keer denk ik al dat 'k er heb gegeten
one time think I already that I there have eaten
"One time I do think I ate there"

For WH-questions and relative clauses, SPOD provides information on the number of cases where the extraction is local (common) or non-local (very rare in Dutch, see for a corpus study Bouma (2017)). The rare case does not occur in Wablieft. In Lassy Small and CGN we find a few. Schippers and Hoeksema (2021) show that long extraction in relative clauses has almost disappeared in Dutch, with the important exception of free relatives. The examples below (from Lassy Small) may serve to illustrate.

(9) Cameron vaarde blind in wat hij dacht dat de juiste richting was
Cameron navigated blindly in what he thought that the correct direction was
"Cameron navigated blindly into what he thought was the right direction"

(10) Ze keek alleen naar wat ze dacht dat het Amerikaanse belang was
she looked only at what she thought that the American interest was
"She only looked to what she thought was the American interest"

3.3 Phrases

SPOD provides counts and average length of noun phrases, prepositional phrases, adjectival phrases and adverbial phrases. For prepositional phrases, further information is provided on their role (modifier of a noun, adjective or verb, prepositional complement, locative-directional complement) and internal structure (with or without "+R"-pronoun). Also, the number of complex prepositions is provided.

If we apply the profiler to the Wablieft corpus, we learn that there are slightly more PP's modifying nouns than verbs.[2] Also, the average length of PP's modifying nouns is slightly smaller. The same trend is observed for other written corpora, but in CGN, PP's more frequently attach to verbs than to nouns, and we do not find a difference in the average length of PP's in these two conditions.

In Dutch, a typical prepositional phrase consists of a preposition directly followed by a noun phrase. However, in case the noun phrase is a pronoun, that pronoun should be a +R pronoun ("er", "daar", "hier", ...), and, moreover, the R-pronoun should precede the preposition (which is then strictly speaking a postposition). These cases may be compared to English *therefore, hereafter, wherein*. The R-pronoun does not have to be adjacent to the postposition, as long as it appears to the left of it. To complicate matters even more, a few prepositions ("tot", "met") have special postpositional variants ("toe", "mee") which have to be used.

(11) De kok smeert **er** de eieren **mee** in (Wablieft)
the cook smears there the eggs with in
"The cook covers the eggs with it"

(12) **Daar** schrok ik zelf even **van** (Wablieft)
there startled I self briefly from
"That gave me a start for a moment"

It is perhaps somewhat surprising to find that this construction is almost equally frequent in the simplified Dutch of Wablieft (1.20%) and Lassy Small (1.47%), even if simpler sentences are readily available, in which R-pronoun and postposition form a unit. As a reviewer notes, it is indeed unclear whether the sentences in which the PP is split are simpler: the non-adjacent variants appear to be more colloquial than the adjacent ones.

(13) De kok smeert de eieren ermee in

(14) Daarvan schrok ik zelf even

3.4 Coordination

Coordinate conjunctions are listed in SPOD according to the number of coordinators (0, 1, 2 or more), the number of conjuncts, the coordinator (*en* "and", *of* "or", *maar* "but", ...), and the category of the conjuncts. This will tell us, for example, that Wablieft has about 5 times more coordinations with *en* than *of*, which in turn is more common than *maar*.

3.5 Comparatives

SPOD provides counts on the number of words that occur with a comparative complement. SPOD distinguishes between the governing word (a comparative adjective, or the words "zo", "even",

2. This result should take into account the observation that in some cases a PP attached to a verb is annotated as "PC" (prepositional complement) or "LD" (locative or directional complement). The distinction between "MOD" and "LD" is in some cases open for discussion.

"meer", "minder", "niet", "niets", "ander", "anders". In addition, results are broken down depending on the category of the complement (NP, PP, VP or S, ADJ or ADV).

In addition, SPOD provides the frequency of correlative comparatives. This construction is illustrated by the following examples.

(15) Hoe meer mensen werken, hoe beter (Wablieft)
how more people work how better
"The more people work, the better"

(16) Hoe groter de misdaad, hoe zwaarder de straf (Wablieft)
how larger the crime how heavier the punishment
"The larger the crime, the heavier the punishment"

3.6 Embedding

SPOD provides information on the depth of embedding of sentences, by providing frequencies of sentences in which finite subordinate sentences are embedded. Embeddings up to a depth of 8 are provided. In Wablieft, there are 19 sentences of depth 3: a finite subordinate sentence is embedded in a finite sentence which is embedded in a finite sentence which is embedded in a finite sentence. An example is:

(17) Ik denk dat ik wel weet waarom we zo weinig gokkers zien die gokken op het internet
I think that I AFF know why we so few gamblers see that gamble on the internet
"I think I know why we see so few gamblers that gamble on the internet"

In Lassy Small, a few cases of depth 4 are encountered, and in the student essays (see section 4 below) even one of depth 5. Depth of embedding is one of the features we expect to see develop over time in the writing of children and adolescents (compare Sampson (2013) for English).

3.7 Parser

For automatically parsed corpora, SPOD uses the meta-data of the parser to provide information on the difficulties encountered by the parser. One aspect is provided by the unknown words, discussed earlier. SPOD also lists the number of cases in which the parser was able to construct a single parse over the full input. Of course, this is only a weak notion of "parser success" since it is not known if the parse is actually correct. Yet, this may indicate the level of surprise encountered by the parser for a particular corpus. For Wablieft, 97.95% of the sentences received a full parse, whereas this number is much lower for the Eindhoven corpus, 92.34%, indicating that at least for the parser, Wablieft is indeed much simpler.

4. Implementation issues

4.1 Running all queries

If a user operates SPOD, she/he selects a corpus, and then selects from a web page with radio buttons the sections and queries that she/he is interested in. The web page also provides the facility to remember the subset of selected queries for later usage. The final selection that a user has to make is the output format. The choice is between a human readable web page (HTML) or a TAB-separated text format that is useful for further automatic processing.

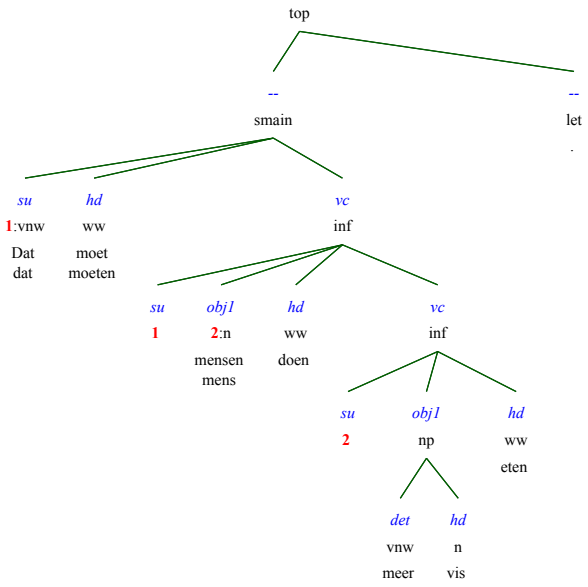

Figure 1: Dependency structure for *Dat moet mensen meer vis doen eten*

Once the choices are selected, the queries have to be evaluated. Some of these queries can take considerable time, in particular of course for the larger corpora. Even so, the resulting page is displayed immediately, but question marks are used in the output for results that are not yet available. Reloading the page updates the page with all results that are available at that point. This appears to be a nice compromise between interactivity of the application, and completeness of the results.

The result of each query and each corpus is cached. In practice, this implies that for the standard corpora of SPOD all results are available instantaneously. If you provide your own corpora, initially your result page may contain many question marks, but the use of caching is very practical in this scenario too, since often the same query is repeated in later sessions as there is no need for the user to keep track of all query results.

4.2 Implementing Queries by means of XPath

The syntactic queries of SPOD are implemented by XPath 2.0 queries. XPath is a standard query language for XML documents. XPath uses path expressions to select nodes or node-sets in an XML document. Since the syntactic analyses of the various treebanks are provided in XML, XPath is a natural choice. In van Noord et al. (2013), the use of XPath for querying the treebanks is motivated and explained in more detail.

The XPath queries required for the syntactic properties of SPOD range from almost trivial to rather complicated. For the complicated queries, the macro system provided by PaQu is used extensively. The use of macro's allows for a notation which is easier to read and comprehend. And furthermore, many macro's are defined which are used in many different queries, leading to better generalization.

As an example, consider the query for the Dutch cross-serial verb cluster construction. A simple example of the construction is given in the following example, with the corresponding dependency structure illustrated in figure 1.

(18) Dat moet mensen meer vis doen eten (Wablieft)

In short, this construction occurs if the direct object in a infinitival verb cluster constituent also plays the role of the subject in a VC complement of that infinitival verb cluster. The head of the dominating VP is typically one of the verbs "zien", "horen", "laten" or "doen". The XPath query which identifies these cases simply states:

```
//node[%PQ_cross_serial_verbcluster_node%]
```

A macro is written between %. This macro is defined as follows, indicating three conditions: the VP occurs in a verb-cluster, the VP is infinitive, and the subject of the VP is co-indexed with the direct object of the verb governing this VP:

```
PQ_cross_serial_verbcluster_node = """
    %PQ_dep_node_in_verbcluster%
    and
    @cat="inf"
    and
    ../node[@rel="obj1"]/%PQ_i% = node[@rel="su"]/%PQ_i% """
```

The further details of these conditions are not specified here further, but the fact that the macro system make the queries easier to understand and use is important, because often the result of SPOD leads to further investigations for a particular construction. For instance, a natural follow up question concerning cross-serial verbs is about the possibilities that more of such verbs co-occur. A query for the combined case then simply looks like:

```
//node[%PQ_cross_serial_verbcluster_node% and
       node[%PQ_cross_serial_verbcluster_node%] ]
```

Despite the attention in linguistic literature to cross-serial verb clusters, and the potential of essentially unlimited sequences of them, even a combination of two such verbs is disappointingly rare.

4.3 Tools for Further Analysis

On the basis of the results provided by SPOD, users often want to zoom in on certain aspects of those results. This is straightforward, since SPOD is integrated with PaQu. In the results page, every result is associated with a direct link to the PaQu page which lists the individual matches. From that page, you have the tools provided by PaQu: visualization of the dependency structures, and the option to obtain counts for a specified set of attributes of the matching nodes. For instance, if you investigate the properties of a corpus and you find an unexpected number of comparative adjectives combining with a comparative complement, then one click will take you to the page with the sentences containing all hits. One further step could then provide you with the frequency overview of all lemma's and postags of these hits.

One further piece of functionality provided by the SPOD results page, is that the results not only provide the frequency counts, but also the average length of the matches (in terms of words). For instance, if you are interested in coordinate conjunction, it may be relevant to see that coordinate conjunctions with a single coordinator differ in length with respect to the choice of that coordinator. Typically, a coordination with "of" ("or" in English) is shorter than a coordination with "en" ("and"). If you click on an average length of a syntactic property, then a graph displaying the full distribution is provided as well. An example is given in figure 2.

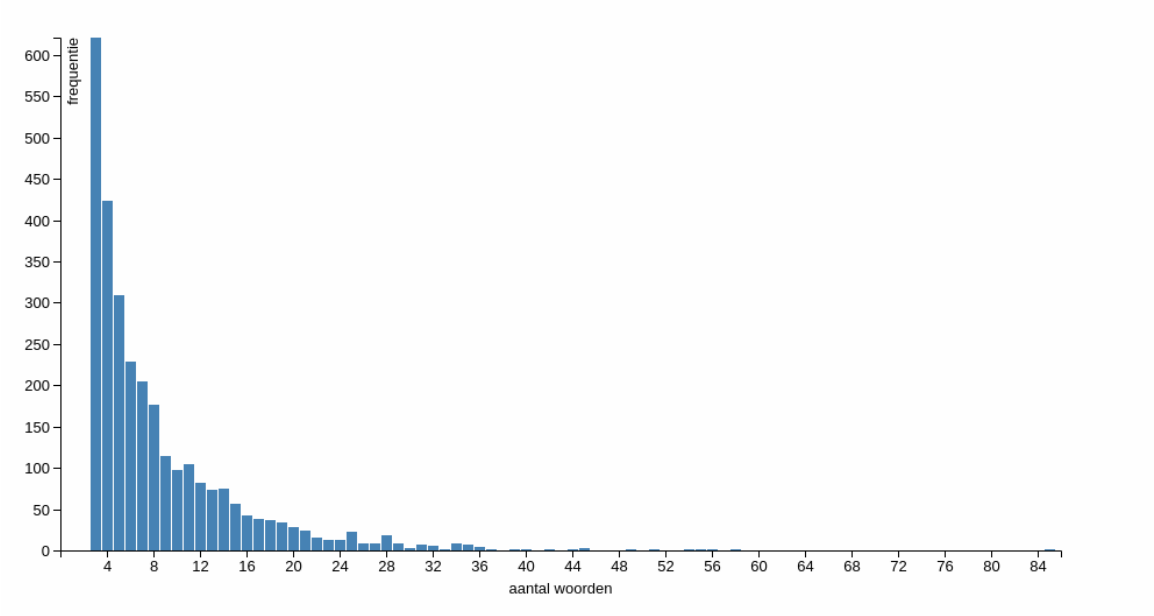

Figure 2: Screenshot of SPOD, displaying the distribution of lengths of coordinate conjunctions with coordinator "of" ("or").

5. Case Study: Phrasal categories in learner corpora

5.1 Goal of the case study

For this case study, we made use of several corpora to study some aspects of the development of writing in Dutch children and young adults. In particular, we want to illustrate the usefulness of PaQu and SPOD for quickly finding developmental trends by applying them to a set of parsed learner corpora, to be described below. Our focus in this case study will be the prevalence of four types of phrases in these corpora, developments in their average size, and for one of the phrasal categories, prepositional phrases, we also look at developments among their grammatical subtypes.

5.2 The learner corpora

The Basiscript corpus of elementary school essays (Tellings et al. 2018), henceforth ES, is a 9 million word corpus. It contains texts by children from grade 4 to grade 8. HS, a corpus of high school essays made available by Kees de Glopper is 314K words in size and has texts from grade 1, 3, 4 and 5. Grade 2 and grade 6 texts were not collected. The development of this corpus was made possible by a grant from the Kennisrotonde van het Nationaal Regieorgaan Onderwijsonderzoek, a Dutch funding agency for education-oriented research. STUD is a corpus of student essays (175K words), all from humanities students, compiled by Jack Hoeksema. LING is a corpus of Dutch linguistics articles (100K words), and added to the above list for the purpose of comparison. Since we were interested in syntactic structures, the STUD and LING corpora were depleted of tables, lists, references, and example sentences. We use CGN, the corpus of spoken Dutch, and the Eindhoven corpus (623K words), a corpus of mainly newspaper texts for purposes of comparison. The high-school texts are from pupils with the school types HAVO and VWO, which correspond to different levels of education, VWO being the highest and HAVO the second highest in terms of academic standards. In the following, we will mostly ignore school type information.

	CGN	ES	HS	STUD	LING	Eindhoven
NP	67.2	97.1	97.5	99.6	99.4	97.8
PP	31.2	43.8	68.7	86.9	88.2	70.3
AP	31.4	44.7	64.9	78.2	82.8	62.6
AdvP	47.0	52.1	60.9	70.2	73.5	58.6
MLU	7.8	9.7	15.2	20.3	23.8	15.4

Table 1: Phrasal occurrences (in pct) in six corpora and mean length of utterance (MLU)

5.3 Phrasal categories

SPOD searches for the four main categories of Noun Phrases (NP), Prepositional Phrases (PP), Adjective Phrases (AP) and Adverb Phrases (AdvP), and provides a listing of the number of such phrases found, the percentage of sentences with such a phrase and the average length of these phrases. NPs may contain or consist of common nouns, proper names, or pronouns. The sentence *You are your own worst enemy* hence contains two NPs, one of minimal length (*you*) and one of length 4 (*your own worst enemy*).

In spoken languages, utterances without NPs are common. Often, they are simple responses in a dialogue, such as (in English) *OK!, Yes! Oh well!, Right.* In written language (unless it is a dialogue such as in Whatsapp interactions) the percentage of sentences without NPs drops to 3 pct or less. The same is not true for PPs , APs and AdvPs. They show considerable variation in our written corpora.

Based on earlier work on English academic writing styles (Biber and Gray 2010, Biber and Gray 2016), we expect to find more prepositional phrases and longer NPs in academic texts (represented by the STUD and LING corpora) than in spoken Dutch and newspaper Dutch. The elementary school and high-school corpora are expected to show a shift toward the above-mentioned academic features, assuming that these results for English generalize to Dutch (and assuming that Dutch linguistic writing is representative for academic writing style).

5.4 Findings

In Table 1, we list the percentages, for our corpora, of utterances with occurrences of each of the four types of phrases. As already mentioned, the percentages for NP are very similar and at ceiling level, except for the spoken Dutch corpus CGN. For the three remaining types of phrases, we see relatively low percentages for spoken Dutch, and rising numbers from ES to STUD and LING. The newspaper texts from the Eindhoven corpus show percentages similar to the HS corpus, and remain below those for the academic registers of the STUD and LING corpora. The high percentages for PP and AP are in line with the findings for English of Biber and his associates (Biber and Gray 2010, Biber and Gray 2016, Staples et al. 2016). In the last row of Table 1 we put the mean length of utterances (MLU) for the various corpora.

We should note that the higher percentages for HS and STUD compared to ES may in part be attributed to the fact that the latter corpora have longer sentences. The longer the sentences, the more likely they are to contain a phrase of a given type. Note that the same is not true for the length of phrases. Longer sentences do not necessarily contain longer phrases.

This point can be made forcefully if we compare the learner corpora with the data from the Wablieft corpus of simple Dutch. In Table 2, we present data on the average length of phrases in the corpora listed in Table 1, plus the Wablieft corpus. Note that Wablieft has an MLU of 8.1, well below that for elementary school essays, and similar to that of spoken Dutch. However, the average length of NP and PP in Wablieft is higher than in our elementary school corpus Basiscript.

We note that the average size of adverbial phrases is the same for all corpora and near bottom, the theoretical minimum being 1.0 words per phrase. We also notice relatively little variation in size

	CGN	ES	HS	STUD	LING	Eindhoven	Wablieft
NP	2.1	1.8	2.5	3.2	3.6	3.2	2.3
PP	3.7	3.2	4.1	4.8	5.4	4.7	3.4
AP	1.4	1.5	1.6	1.6	1.5	1.6	1.4
AdvP	1.1	1.1	1.1	1.1	1.1	1.1	1.1

Table 2: Mean length (in words) of 4 types of phrases in 7 corpora

	Pred	N-mod	Adv	PP-compl
ES4	19	645	1233	913
ES5	19	750	1308	819
ES6	28	952	1458	681
ES7	40	1049	1810	1026
ES8	39	1124	1959	1110
HS1	66	2444	2652	1699
HS3	72	2887	3042	2188
HS4	111	3207	3318	2418
HS5	132	4020	3774	2607
STUD	164	5348	4884	3016

Table 3: Developments in four types of PP (normalized counts per 10 thousand utterances)

among APs and considerable variation among NPs and PPs. These two categories grow continually in size from elementary school to university, and peak in the works of professional linguists. It would seem, therefore, that the average size of these two categories may be a good estimator of how far pupils are on the road from a completely oral child language to a full-fledged academic register. We also want to note that the fact that some categories (AdvP in particular) show no growth in size over the entire developmental period, whereas others almost double in size, is a finding we had no reason to expect.

The data presented above do not distinguish between the various kinds of prepositional phrases, in particular adverbial modifiers, PP complements to verbs, nouns and adjectives, and PP modifiers of nouns and PP predicates (either copula constructions or secondary predicates). However, SPOD can distinguish among these cases. In Table 3, we present data for 5 grades of elementary school (ES 4 to 8), four grades of high school (HS1, HS3, HS4 and HS5) and university students (STUD). Raw numbers would not be terribly revealing given the differences in size of the various corpora, so the numbers in the table represent occurrences per 10,000 sentences. For all of these categories there is a substantial jump for ES8 to HS1 that is larger than for e.g. ES7 to ES8. This is attributable to the fact that the high-school essays are not representative of all pupils but only of the higher levels (HAVO, VWO). More interesting for us is the fact that almost all columns show continuous increases in numbers all the way from early elementary school to university. PP-complements form a partial exception to this pattern, as they drop a bit after grade 4, before they start a prolonged rise toward the end of elementary, throughout high-school and university. We note that this category is somewhat problematic due to a lack of agreement among linguists on what counts as a PP-complement to a verb. Cases such as *met de studenten praten* 'talk with the students' have been analyzed by some (Broekhuis 2004) as involving an adverbial modifier, on a par with *met de studenten zwemmen* 'swim with the students'. Others (Schermer-Vermeer 2006, Vandeweghe 2011, Hoeksema and Napoli 2019) argue that *met de studenten* is an argument of *praten*, not a modifier. The issue depends on which tests for argumenthood one considers to be decisive, and has not been resolved. PaQu treats *met*-PPs in combinations with verbs of communication as modifiers.

Among the two most common types of PP, in particular PP-modifiers of nouns and adverbial PPs, growth in the former category outpaces that of the latter, as the following graph in figure 3,

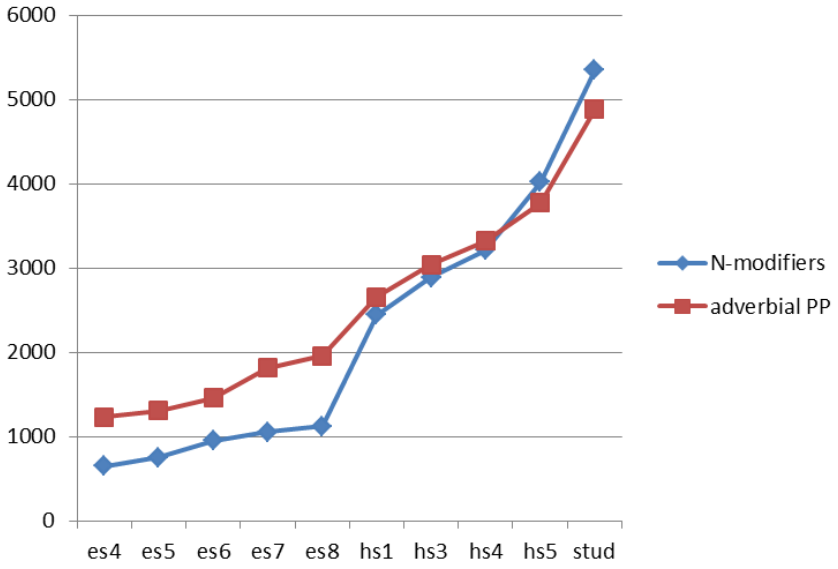

Figure 3: Usage of two types of PP, by school level

based on the two middle columns of Table 3, illustrates. This result appears to confirm the special status of NP modification in the academic register noted by Biber and Gray (2010).

5.5 Conclusions of the case study

Noun phrases and prepositional phrases show a double increase along the path from 4th grade elementary school to university level. They increase in prevalence (especially PPs) as well as in their average size. This fits nicely with findings for English academic writing in comparison to other types of usage by Biber and associates, but also with that of Heylighen and Dewaele (2002), who found that more formal registers, such as academic writing, were characterized by higher levels of nouns and prepositions.

6. Concluding Remarks

SPOD, the Syntactic Profiler of Dutch, is a tool for automatically generating a large set of syntactic properties from a corpus by means of a structured collection of queries. We discussed the syntactic properties that are available in SPOD. Furthermore, we showed how SPOD interacts with PaQu, to enable further in-depth study of the relevant syntactic constructions, and to critically assess the raw numbers provided by SPOD. To illustrate the power of SPOD, we presented a case study on the development of the use of the major syntactic categories NP, PP, AP and AdvP. The study of this development in essence boils down to running SPOD over a number of corpora from the relevant age groups, and compare the results. In this way, SPOD enables and greatly facilitates the corpus study of syntactic phenomena.

Acknowledgements

The development of SPOD has been funded by the Dutch national CLARIN project Common Lab Research Infrastructure for the Arts and Humanities, CLARIAH.

References

Augustinus, Liesbeth, Vincent Vandeghinste, and Frank Van Eynde (2012), Example-based treebank querying, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, pp. 3161–3167.

Biber, Douglas (1993), Using register-diversified corpora for general language studies, *Computational Linguistics* **19** (2), pp. 219–241, MIT Press.

Biber, Douglas and Bethany Gray (2010), Challenging stereotypes about academic writing: Complexity, elaboration, explicitness, *Journal of English for Academic Purposes* **9**, pp. 2–20, Elsevier.

Biber, Douglas and Bethany Gray (2016), Phrasal versus clausal discourse styles: A synchronic grammatical description of academic writing contrasted with other registers, *in* Biber, Douglas and Bethany Gray, editors, *Grammatical Complexity in Academic English: Linguistic Change in Writing*, Cambridge University Press, pp. 67–124.

Bloem, Jelke (2020), Een corpus waar alle constructies in gevonden zouden moeten kunnen worden? corpusonderzoek met behulp van automatisch gegenereerde syntactische annotatie., *Nederlandse Taalkunde*, Amsterdam University Press.

Bouma, Gosse (2017), Finding long-distance dependencies in the Lassy corpus, *in* Hilke Reckman, Maarten Hijzelendoorn, Lisa Lai-Shen Cheng and Rint Sybesma, editors, *Crossroads Semantics: Computation, experiment and grammar*, Benjamins.

Bouma, Gosse, Gertjan van Noord, and Robert Malouf (2001), Alpino: Wide-coverage computational analysis of Dutch, *in* Daelemans, Walter, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in The Netherlands 2000*, Rodopi.

Broekhuis, Hans (2004), Het voorzetselvoorwerp, *Nederlandse Taalkunde* **9** (2), pp. 97–131, Van Gorcum.

de Sutter, Gert, Dirk Speelman, and Dirk Geeraerts (2005), Regionale en stilistische effecten op de woordvolgorde in werkwoordelijke eindgroepen, *Nederlandse Taalkunde* **10**, pp. 97–128, Amsterdam University Press.

Heylighen, Francis and Jean-Marc Dewaele (2002), Variation in the contextuality of language: An empirical measure, *Foundations of science* **7**, pp. 293–340, Springer.

Hoeksema, Jack and Donna Jo Napoli (2019), Degree resultatives as second-order constructions, *Journal of Germanic Linguistics* **31**, pp. 225–297, Cambridge University Press.

Huybregts, Riny (1984), The weak inadequacy of context-free phrase structure grammars, *in* de Haan, Ger, Mieke Trommelen, and Wim Zonneveld, editors, *Van Periferie naar Kern*, Foris, pp. 81–99.

Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong (2013), From high heels to weed attics: a syntactic investigation of chick lit and literature, *Proceedings of the Workshop on Computational Linguistics for Literature*, Atlanta, Georgia, pp. 72–81.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993), Building a large annotated corpus of english: The Penn treebank, *Computational Linguistics*, MIT Press.

Moortgat, Michael, Ineke Schuurman, and Ton van der Wouden (2000), CGN syntactische annotatie. internal report Corpus Gesproken Nederlands.

Odijk, Jan (2015), Linguistic research with PaQu, *Computational Linguistics in the Netherlands Journal* **5**, pp. 3–14, CLIN.

Odijk, Jan (2020), De verleidingen en gevaren van GrETEL, *Nederlandse Taalkunde* **25** (1), pp. 7–37, Amsterdam University Press.

Odijk, Jan, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang (2017), The parse and query (PaQu) application, *in* Odijk, Jan and Arjan van Hessen, editors, *Clarin in the low countries*, Ubiquity Press, London.

Pander Maat, Henk, Rogier Kraf, APJ van den Bosch, Nick Dekker, M van Gompel, S de Kleijn, Ted Sanders, and K van der Sloot (2014), T-Scan: a new tool for analyzing Dutch text, *Computational Linguistics in the Netherlands Journal* **4**, pp. 53–74. https://clinjournal.org/clinj/article/view/40.

Pauwels, Anita (1953), *De plaats van hulpwerkwoord, verleden deelwoord en infinitief in de Nederlandse bijzin*, Koninklijke Commissie voor Toponymie en Dialectologie, Leuven.

Roland, Douglas, Frederic Dick, and Jeffrey L. Elman (2007), Frequency of basic English grammatical structures: A corpus analysis, *Journal of Memory and Language* **57** (3), pp. 348–379, Elsevier.

Sampson, Geoffrey (2013), The structure of children's writing, *in* Sampson, Geoffrey and Anna Babarczy, editors, *Grammar without Grammaticality: Growth and Limits of Grammatical Precision*, Walter de Gruyter, pp. 155–171.

Schermer-Vermeer, Ina (2006), Worstelen met het voorzetselvoorwerp, *Nederlandse Taalkunde* **11** (2), pp. 146–167, Van Gorcum.

Schippers, Ankelien and Jack Hoeksema (2021), Langeafstandsverplaatsing in het Nederlands, Engels en Duits: de sandwich ontleed, *Nederlandse Taalkunde*, Amsterdam University Press.

Schuurman, I., M. Schouppe, T. Van der Wouden, and H. Hoekstra (2003), Cgn, an annotated corpus of Spoken Dutch, *in* Abbeilé, A., S. Hansen-Schirra, and H. Uszkoreit, editors, *Proceedings of 4th International Workshop on Language Resources and Evaluation*, Budapest, pp. 340–347.

Staples, Shelley, Jesse Egbert, Douglas Biber, and Bethany Gray (2016), Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre, *Written Communication* **33**, pp. 149–183, Sage.

Tellings, Agnes, Nelleke Oostdijk, Iris Monster, Franc Grootjen, and Antal Van Den Bosch (2018), Basiscript: A corpus of contemporary dutch texts written by primary school children, *International Journal of Corpus Linguistics* **23**, pp. 494–508, John Benjamins.

uit den Boogaart, P. C. (1975), *Woordfrequenties in geschreven en gesproken Nederlands*, Oosthoek, Scheltema & Holkema, Utrecht. Werkgroep Frequentie-onderzoek van het Nederlands.

van der Beek, Leonoor, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada (2002a), Algorithms for linguistic processing. nwo pionier progress report. www.let.rug.nl/vannoord/alp/midterm.pdf.

van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Gertjan van Noord (2002b), The Alpino dependency treebank, *in* Mariët Theune, Hendri Hondorp, Anton Nijholt, editor, *Computational Linguistics in the Netherlands 2001*, Rodopi.

van der Wouden, Ton, Gosse Bouma, Marjo van Koppen, Frank Landsbergen, Jan Odijk, and Matje van de Camp (2015), Enriching a descriptive grammar with treebank queries, *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 13 – 25.

van Eynde, Frank (2005), Part of speech tagging en lemmatizering van het D-COI corpus. http://www.let.rug.nl/vannoord/Lassy/POS_manual.pdf.

van Eynde, Frank, Jakub Zavrel, and Walter Daelemans (2000), Part of speech tagging and lemmatisation for the spoken Dutch corpus, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, European Language Resources Association (ELRA), Athens, Greece. http://www.lrec-conf.org/proceedings/lrec2000/pdf/216.pdf.

van Noord, Gertjan (2006), **A**t **L**ast **P**arsing **I**s **N**ow **O**perational, *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, Leuven, pp. 20–42.

van Noord, Gertjan (2009), Huge parsed corpora in Lassy, *in* van Eynde, Frank, Anette Frank, Koenraad de Smedt, and Gertjan van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, LOT Occasional Series, Groningen, The Netherlands.

van Noord, Gertjan and Gosse Bouma (2009), Parsed corpora for linguistics, *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, ILCL '09, Association for Computational Linguistics, USA, p. 33–39.

van Noord, Gertjan, 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* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 147–164.

van Noord, Gertjan, Ineke Schuurman, and Gosse Bouma (2019), Lassy syntactische annotatie. http://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf.

Vandeghinste, Vincent, Bram Bulté, and Liesbeth Augustinus (2019), Wablieft: An easy-to-readnewspaper corpus for dutch,proceedings of the clarin annual conference, *Proceedings of the CLARIN Annual Conference*, Leipzig, Germany, pp. 188–191.

Vandeweghe, Willy (2011), Het voorzetselvoorwerp en de hiërarchie der objecten, *Nederlandse Taalkunde* **16** (1), pp. 88–101, Amsterdam University Press.