Gosse Bouma

Corpus-Evidence for True Long-Distance Dependencies in Dutch

Abstract Long-distance dependencies have been studied extensively in syntactic theory. Yet, true long-distance dependencies, spanning more than a single predicate, appear to be rare in actual use. In this paper, we present the results of searching for such dependencies in a large, automatically annotated, treebank for Dutch, concentrating on phenomena that have recently been subject to debate, and where conflicting claims have been made regarding their productivity and existence.
Our results suggest that in Dutch, true long-distance dependencies are rare and have limited productivity. We also show that a popular strategy for avoiding such dependencies, resumptive prolepsis, is much more frequent and productive. Finally, we demonstrate that the annotation also facilitates searching for parasitic gaps, even though the construction itself is outside the scope of the computational grammar.

Keywords Long-distance dependencies, corpora, Dutch, resumptive prolepsis, parasitic gaps

1 Introduction

While syntactic theory has highlighted the possibility of potentially unbounded dependencies in WH-questions and relative clauses, in actual language use the dependencies introduced by a WH-question or relative clause are often very short and rarely span more than a single clause. To what extent genuine long-distance dependencies occur in natural language is therefore still an open question. Corpus-based research into this issue has been hindered by the fact that long-distance dependencies are difficult to find using search patterns consisting of lexical items and/or part-of-speech tags only. Syntactically annotated treebanks are more promising, as in theory they offer the kind of annotation required to identify long-distance dependencies. The Penn Treebank (Marcus et al. 1994) for instance, explicitly marks the relationship between WH-phrases

and relative pronouns and the 'extraction' site. However, carefully annotated and manually corrected treebanks are limited in size, while making claims about the possibility and productivity of certain long-distance dependencies requires corpora of considerable size. The alternative that we opt for in this paper is to work with automatically annotated data. The Alpino parser for Dutch (van Noord 2006) uses a linguistically motivated grammar and achieves high coverage and precision on most text genres.[1] The parser has been used to create the Lassy Large (van Noord et al. 2013), a large syntactically annotated corpus.

In this paper, we present the results of searching for four kinds of long-distance dependencies in an automatically annotated treebank for Dutch. We concentrate on phenomena that have recently been subject to debate, and where conflicting claims have been made regarding the question whether these constructions actually occur with some frequency in spontaneous language use. In particular, we will provide an answer to the following questions:

— To what extent do we find collocational effects in WH-questions and relative clauses involving a true long-distance dependency (Verhagen 2006)?
— To what extent do we find long-distance dependencies into infinitival clauses introduced by the optional complementizer *om*?
— What is the relationship between resumptive prolepsis (Hoeksema and Schippers 2012) and (the absence of) non-local dependencies?
— To what extent do we find parasitic gap constructions involving R-pronouns (Everaert et al. 2015) in actual text?

2 Background

One of the central topics in theoretical syntax is the proper analysis of non-local dependencies of the kind found in WH-questions and relative clauses. Rather different solutions have been proposed in various theoretical frameworks (among others in Transformational Grammar [Chomsky 1977], Categorial Grammar [Morrill 1995; Steedman 2000], GPSG [Gazdar et al. 1985], HPSG [Bouma et al. 2001], and LFG [Kaplan and Zaenen 1989]). One of the surprising facts is that there is still considerable disagreement about what the relevant data are and whether these are to be accounted for in syntax or by an appeal to general

1 In a recent comparison using the Universal Dependencies Lassy Small Corpus (http://universaldependencies.org/#nl_lassysmall), Alpino achieved labelled accuracy scores that were 4–7% higher than three state-of-the-art dependency parsers (including SyntaxNet) (Bouma and van Noord 2017).

cognitive constraints (Hofmeister and Sag 2010). Another observation that is somewhat at odds with the claims of most studies in theoretical syntax is that in actual usage, sentences involving a true long-distance dependency are rare, and often involve the same matrix verb and subject, suggesting that these are all variants of a small set of constructions (Verhagen 2006).

A corpus study can help to provide more insight in the frequency with which certain long-distance dependency constructions occur, and the amount of variation observed with each phenomenon. While WH-questions and especially relative clauses occur with some frequency in most corpora, cases that involve a true long-distance dependency (i.e. cases where the 'gap' is located in a subordinate clause) are not very frequent, and thus we will concentrate on material obtained from a large, but automatically parsed, corpus. This raises the question how accurate our results will be.

In computational linguistics, it has been observed that while statistical parsers now achieve very acceptable accuracies in general, this is not always the case when concentrating on more challenging aspects of syntax, such as properly accounting for non-local dependencies (Rimell et al. 2009; Candito and Seddah 2012). As we are using a corpus that was automatically annotated using the Alpino parser (van Noord 2006), this study can also give some insights into the accuracy of Alpino into analyzing non-local dependencies.

3 Non-local dependencies in the Lassy Corpus

The Lassy Large corpus (van Noord et al. 2013) is a corpus of contemporary Dutch that has been annotated with syntactic information. Annotation consists of lemmas, part-of-speech tags, constituent structure and dependency relations. It is composed of all material in the SONAR500 corpus (a mixed corpus of Dutch, containing texts from 18 different genres, i.e. adminstrative, autocues, magazines, legal, proceedings, web, etc., 41M sentences) (Oostdijk et al. 2013), Dutch Wikipedia (2011 dump, 9M sentences), EMEA (European Medicines Agency, 1M sentences), EUROPARL (proceedings of the European Parliament, 1M sentences), and various smaller sources. Syntactic annotation was done automatically using the Alpino parser (van Noord 2006). A small part of the corpus has been manually verified (Lassy Small, 65K sentences). Lassy Small and the Wikipedia-part of Lassy Large can be explored online.[2] In the examples below (Figure 1), we formulate queries using XPATH, as documented in Odijk (2015) and Augustinus et al. (2017).

2 http://zardoz.service.rug.nl:8067/

voor wie is de hulp bedoeld?
For who is the aid meant
'Who is the aid meant for?'

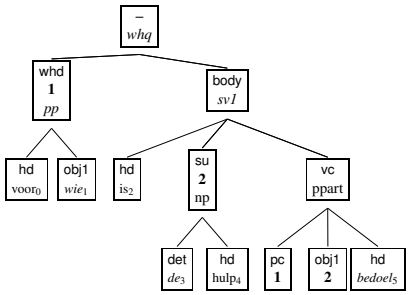


Figure 1: WH-question and corresponding syntactic dependency tree.

In this paper, we will be mostly concerned with syntactic constituency and dependency relations. As an example, consider the annotation of the WH-question sentence in Figure 1. The sentence initial WH-constituent *voor wie* is labeled with category PP. Internally, it consists of a head and a dependent labeled with the dependency relation `obj1` (used for objects of verbs and prepositions). The clause itself is a passive, headed by the auxiliary *is*, and containing two dependents: a subject and a verbal complement headed by a passive participle *(bedoeld)*. The passive participle phrase contains two empty nodes: a prepositional complement node co-indexed with the fronted PP and an object node co-indexed with the subject. The co-indexing between the initial PP and the prepositional complement of *bedoeld* expresses a non-local dependency. Following standard linguistic practice, we will sometimes refer to the latter type of node as a *'gap'*, even though the HPSG formalism on which the Alpino grammar is based does not actually employ gaps in its analysis of non-local dependencies.

Syntactically annotated corpora are useful for obtaining information about the distribution of such dependencies in actual usage. As a first example of how one can use a corpus to study non-local dependencies, we will look at the distribution of gaps in simple relative clauses. Simple finite clauses consist of a finite verb and one or more dependents that function as subject, direct object, indirect object, prepositional complement, etc. The dark bars in Figure 2 show that while all of these can be relativized, in 77% of the cases the gap is a subject. One might think that this is a consequence of the fact that subjects are simply more frequent than other dependents. The grey bars in Figure 2 show the distribution of all dependents in simple relatives (i.e. gapped or not). Only 37% of all dependents are subjects. This shows that in the vast majority of relative clauses, the gapped element is a subject, and that this preference is not (only) a consequence of the fact that in simple finite clauses, subjects are the most frequent dependents in general.

The statistics for gaps in simple relatives were obtained by running the following query on Lassy Small:

```
(1)  //node[ not(@word or @cat) and
               number(@index) = ../../node[@rel="rhd"]/number(@index)
           ]
```

This query searches for a node that has no `word`- or `cat`-attribute. This guarantees that the node does not correspond to a substring in the input sentence, i.e. it is a 'gap' (Figure 2).

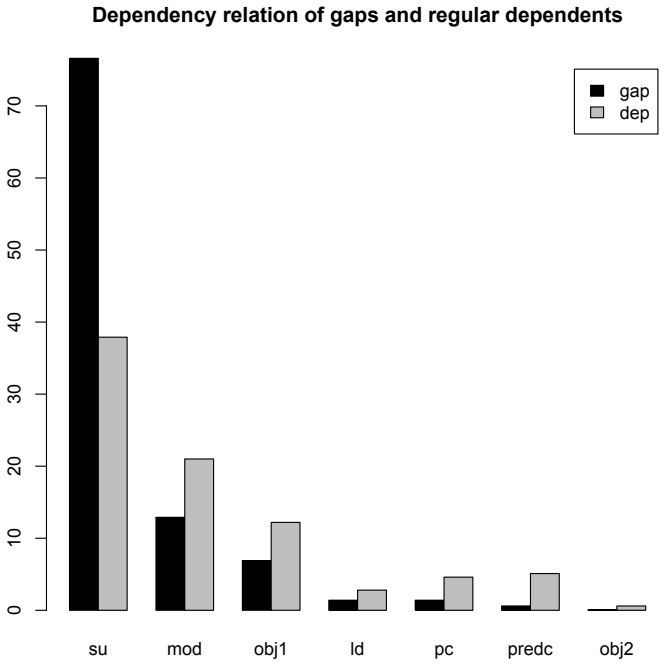


Figure 2: Distribution of dependency labels of gaps and regular dependents in simple relative clauses in Lassy Small.

Next, it requires that its `index` attribute has the same value as the node with dependency label `rhd` (this is the head of a relative clause), that occurs as a daughter (`'/node'`) of the grandmother (.../..) of the node itself. This ensures that we are only looking at 'local' instantiations of long-distance dependencies. It gives rise to over 8,000 hits.

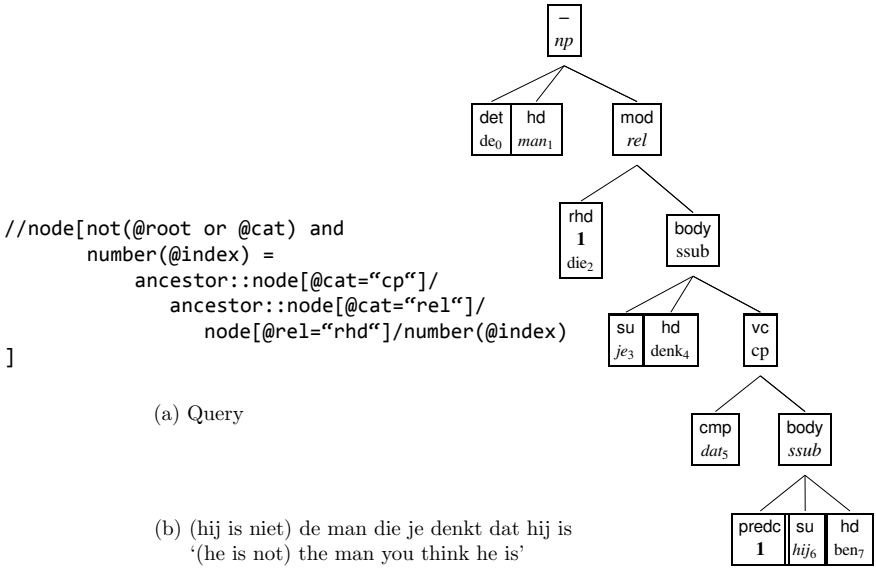
To obtain statistics for all dependents in the same set of relative clauses (the grey bars), we need to formulate a slightly more complex query:

```
(2)  //node[ not(@rel="hd") and
             ../node[ not(@word or @cat) and
                        number(@index) =
                           ../../node[@rel="rhd"]/number(@index)
                      ]
           ]
```

This query matches any non-head node that has a sister that meets the requirements of the previous query. Thus, we are looking at the same set of simple relative clauses as before, but now we can gather statistics for all non-head dependents (i.e. gapped or regular).

4 True long-distance dependencies

The dependency between a relative clause head and its corresponding gap is truly long-distance if the gap is located in a clause that is subordinate to the matrix verb of the relative clause or WH-question (Figure 3).[3]



(a) Query

(b) (hij is niet) de man die je denkt dat hij is
'(he is not) the man you think he is'

Figure 3: Long-distance dependencies in relative clauses.

3 Candito and Seddah (2012) use a slightly more liberal notion of true long-distance dependency that also includes 'gaps' in nominal and adjectival predicative phrases. Although such cases occur in Dutch, they are ignored in the present study.

There has been some discussion as to what extent such long-distance dependencies occur in (contemporary) Dutch, and whether they are limited to a small set of matrix verbs and subjects or not (Verhagen 2006; Hoeksema and Schippers 2012).

To find true LDDS in Lassy Large, we used the query in Figure 3a. It searches for a 'gap' dominated by a finite subordinate clause introduced by a complementizer[4] (i.e. its category is CP, for *complementizer phrase*), which in turn has to be dominated by a relative clause node (or WHQ node in the case of WH-questions). Furthermore, the index of the node has to be identical to the index of the head of the relative clause. An example of such a configuration is given in Figure 3b.

For the complete Lassy Large corpus, the query returned 270 hits for relatives, 73 of these were true LDDS (27 %). The query for WH-questions returned 2,601 hits, of which 344 cases were true LDDS (13 %). The distribution of matrix verbs in these examples is given in Table 1.

Table 1: Counts for matrix verbs in relative clauses and wh-questions with a true ldd.

Verb	N (rel)	N (wh)	Verb	N (rel)	N (wh)
denken ('to think')	52	252	hopen ('to hope')	3	1
willen ('to want')	7	49	weten ('to know')	2	0
zeggen ('to say')	4	5	vermoeden ('to suspect')	2	0
vinden ('to find')	3	33	zien ('to see')	0	1
			wensen ('to wish')	0	1
			verwachten ('to expect')	0	2

The dominance of *denken* is striking, and confirms to some extent the observations in Verhagen (2006).

It should also be noted however, that the corpus contains a fair amount of user generated content from social media. In this text genre, the relative clause *die je/hij/ik/ze denk(t)(en) dat je/hij/ik/ze is/ben (that you think I am* and pronominal variants) is a frequently occurring phrase.

Recently, there has been quite a bit of discussion about the possibility of *weten* as matrix verb in long-distance dependency constructions (Coppen 2013).[5] It has been claimed that only non-factive verbs can be matrix verbs in long-distance dependencies of this kind (Ross 1967). Coppen points out that similar examples involving *weten* can be found relatively easily in literature from the 17th and 18th century, and also suggests that *weten* might not be strictly factive

4 Note that in Dutch, the presence of a complementizer is obligatory in this construction.

5 The discussion in the media was triggered by the phrase *de dag die je wist dat zou komen (the day that you knew that would come)* from a song composed on the occasion of the coronation of King Willem Alexander (2013).

in all contexts. Our results show that even in modern Dutch, the use of *weten* in true LDDS is not completely excluded. These are the two examples with factive matrix verb *weten*:

(3) a. ik ben nog steeds niet de volwassene die ik wist dat ik kon zijn
I am still still not the adult that I knew I could be
'I am still not the grown-up that I knew I could be'

b. ik pak alleen mensen die ik weet da eerlijke kans maken
I grab only people that I know that honest chance make
'I only attack people that I how have an honest chance'

Verhagen (2006) finds that in his corpus (Eindhoven corpus and articles from 'de Volkskrant'), the subject in WH-questions involving a long-distance dependency is almost always a second person pronoun. The distribution in the examples found in true LDDS in the Lassy corpus (Table 2) confirms that this is indeed predominantly the case for WH-questions. For relative clauses, however, a more diverse picture emerges. There is a strong preference for pronominal subjects, but first, second, and third person pronouns are all of approximately the same frequency.

Table 2: Distribution of subjects in matrix clauses in true ldds.

	Relatives	wh-questions
first person	25	9
second person	23	313
third person pronouns	25	11
full NPs	3	13
other	2	

The Alpino grammar specifies lexically which verbs that take a clausal complement can occur as matrix verbs in long-distance dependency constructions (these verbs are sometimes called *'bridge verbs'*). This list is slightly larger than the verbs mentioned in Table 1, and also contains *bedoelen* ('to mean'), *beloven* ('to promise'), and *beweren* ('to claim'). Even if longdistance dependencies are rare, the size of the Lassy Large corpus would lead one to expect that at least for all of these verbs, some examples can be found. Of course, we should keep in mind that the Lassy Large corpus was automatically analyzed and thus some relevant cases may have been missed. For instance, manual inspection of all relatives with matrix verb *beweren* and containing a subordinate clause in the Wikipedia section of Lassy Large did reveal one case involving a long-distance dependency:

(4) de naam waaronder men beweerde dat Menelaos een tempel
the name under-which one claimed that Menelaos a temple
voor Aphrodite had opgericht
for Aphrodite had founded
'the name under which one claims that Menelaos had founded a temple for Aphrodite'

Of all the question sentences with matrix verb *beweer* in Lassy Large (116 cases), not a single one contained a true LDD. Also, manual inspection of all WH-questions with *bedoelen* and *beloven* as matrix verb did not return a single case with a true LDD. It is thus not impossible that examples of true LDDS involving other 'bridge' verbs are present in the corpus, but at the same time these results suggest that they will not be very frequent.

True LDDS are extremely rare in the Lassy Large corpus. For a similar construction in English, relatives involving subject extraction from an embedded clause, Rimell et al. (2009) report that it occurs in 0.4 % of the sentences in their corpora (Wall Street Journal and Brown). The Lassy Large corpus contains more than 50M sentences, and thus even if the recall of the Alpino parser is low on this phenomenon, it seems unlikely that more than several thousand (i.e. 0.002–0.01%) of the sentences in Lassy Large contain a true LDD.

5 Long distance dependencies with non-finite clauses

It is not exactly clear what should be counted as a long-distance dependency. Usually, cases involving an auxiliary or modal as in (5) are not seen as long-distance, even though one might claim that these involve a matrix clause (the auxiliary or modal and the subject) and an embedded non-finite VP.

(5) de kiesdrempel die de partij zelf had ingevoerd
the election-threshold that the party itself had introduced
'the election threshold that the party had introduced itself'

However, there are also verbs that select a *to*-infinitival complement, where the matrix verb cannot be seen as a modal or auxiliary (Cremers 1983). In those cases where the to-infinitival complement is in 'extraposed' position, it can be optionally introduced by the complementizer *om:*

(6) De stichting is verplicht (om) haar winst aan sociale projecten
The foundation is obliged (CMP) her profit to welfare projects
uit te keren
out to turn
'The foundation is obliged to give her profit to welfare projects'

It seems reasonable to categorize relative clauses that involve a dependency with a gap inside a *to*-infinitive of this kind as true LDDs as well. An interesting question in this case is the role of the optional complementizer. The presence or absence of *om* is influenced by various factors involving sentence complexity, such as distance between the matrix verb and complement, frequency of the matrix verb, and frequency with which the matrix verb occurs with a VP-complement (Bouma 2017). Whether the presence of a long-distance dependency also influences the likelihood of the complementizer *om* is unclear. For instance, Bennis (2000) presents example (7-a), where *om* is marked as optionally possible. Broekhuis et al. (1995) present example (7-b), but add in the discussion that *'it must be mentioned that the complementizer is preferably dropped'*.

(7) a. Waar is Jan bang (om) over te praten
Where is John afraid (CMP) over to talk
'What is John afraid of to talk about'
b. Wat heeft Jan geprobeerd om te lezen
What has John tried CMP to read
'What has John tried to read'

We tried to find cases like this in the corpus. The search for cases that are introduced by *om* is relatively straightforward, and requires only a minor variation of the query given above for finite complements (i.e. instead of a node with category CP we now search for the same configuration with a node of category OTI (for *om-te*-infinitive)):

(8) een boek dat je intellect simpelweg weigert om serieus te nemen
a book that your intellect simply refuses CMP seriously to take
'a book that your intellect simply refuses to take seriously'

When searching for cases where the complementizer is absent, we added an additional constraint to the query that requires that the *te*-infinitive contains at least one dependent that follows the matrix verb but precedes the verb heading the infinitival clause, as in (9-a). This ensures that the infinitive is indeed an '*extraposed*' complement, and has not been integrated into the matrix clause as a

result of a process that is known as *'verb raising'*, as in (9-b). In the latter case, it is unclear whether there is indeed a long-distance dependency.

(9) a. organisaties die ik vergeten ben een adreswijziging te sturen
organisations that I forgot am an address-change to send
'organisations to which I forgot to send a change of address'
b. organisaties die ik een adreswijziging ben vergeten te sturen

The results for searching for true LDDS in infinitival complements are given in Table 3. There is quite a bit of variation in matrix verbs in both cases (16 different types for *om-te*-infinitives, and 22 different types for *te*-infinitives). The only verb that occurs with a high frequency (21 hits) is *achten* ('to suppose') in the *te*-infinitive case, as in (10). This is unexpected, as *achten* is not a very frequent verb in general.

Table 3: Counts for true ldds involving infinitival complements

	hits	valid	verb types
om-te-infinitives	81	28	16
te-infinitives	275	75	22

(10) conversaties die ze geacht worden niet te horen
conversations that they supposed are not to hear
'converstations that they were not supposed to hear'

Our results confirm that true LDDS are possible with both *om-te*-infinitives and *te*-infinitives, and that this is possible for a wide range of matrix verbs. The results do not give a clear answer to the question whether true LDDS are less likely if *om* is present, as the two data-sets are not very comparable (i.e. we added an additional constraint to the query for *te*-infinitives).

Manual checking was necessary to obtain the results in this section and the preceding section. As a result, we can observe that the precision of the Alpino parser on true LDDS in relative clauses in Lassy Large is 35% (73/209), 13% (344/2601) for true LDDS in questions, 35% for *om-te*-infinitives (28/81) and 27% for *te*-infinitives in extraposed position. This may not seem very high, but, with the exception of WH-questions, it is in fact comparable to the performance of the best performing system in Rimell et al. (2009) on subject extraction from an embedded clause. It should also be noted that these make up a tiny portion of the corpus as a whole, and thus, the effect on parser accuracy in general is negligible.

6 Resumptive prolepsis

Hoeksema and Schippers (2012) present results from a diachronic corpus study suggesting that true LDDs are in decline in Dutch, and that, especially in relative clauses, they are being replaced by a construction referred to as *'resumptive prolepsis'* by Salzmann (2006) and which involves a relative clause headed by *waarvan* ('of which') or *van wie* ('of whom') and a 'resumptive' pronoun in an embedded clause:

(11) a. 45 mogelijke van Goghs **waarvan** onduidelijk is of **ze**
45 potential van Gogh's of-which unclear is whether they
echt of vals zijn
true or fake are
'45 potential van Gogh's of which it is unclear whether they are true or false'

b. iemand van wie ze denkt dat hij haar man is
somebody of-which she thinks that he her husband is
'somebody that she thinks is her husband'

The Alpino parser does analyse these as relative clauses where the relative head is co-indexed with a gap in the matrix clause that is labeled as a modifier. It does not establish a relation between the pronoun in the subordinate clause and the relative clause head. To find instances of this construction involving the adverbial PP *waarvan*, we used the following query:

```
(12) node[ @cat="rel" and node[@lemma="waarvan"]]/
        node[ .//node[@rel="mod" and @index]]//
            node[ @cat="cp" and (@rel="su" or @rel="vc")]//
                node[ @pt="vnw" and (@rel="su" or @rel="obj1") and
                      (@vwtype="pers" or @vwtype="aanw") ]
```

This query searches for relative clauses headed by *waarvan*, dominating a node that has a descendant that is an indexed modifier (the gap) and which has a descendant that is a finite subordinate clause with dependency label su or vc. The latter constraint ensures that the CP is indeed a complement, and not a modifier. Finally, the subordinate clause has to contain a personal or deictic pronoun with dependency label su (for subject) or obj1 (for direct object, of a verb or preposition). The query for *van wie*-cases is similar except for the definition of the relative clause head.

This query, while only approximating the requirements of the resumptive prolepsis construction, returns more than 9,500 hits and turns out to be quite

Table 4: Distribution of matrix verbs and pronouns in *waarvan/van wie*$_i$*... pronoun*$_i$ constructions

matrix verb	hits	%	pronoun	hits	%
weten ('to know')	1489	15.6	ze	3537	37.1
denken ('to think')	1324	13.9	het	2340	24.6
bekend zijn ('be known')	851	8.9	hij	1282	13.4
zeggen ('to say')	709	7.1	die	752	7.9
vermoeden ('to suspect')	498	5.2	zij	729	7.9
hopen ('to hope')	396	4.2	deze	581	6.1
verwachten ('to expect')	392	4.2	er	280	3.0
vinden ('to .nd')	376	3.8	hem	134	1.4
veronderstellen ('to suppose')	254	2.6	dat	117	1.2
beweren ('to claim')	249	2.6	dit	91	1.0
other	3413	34.3	*other*	1.0	

accurate. In a random sample of 100 sentences, we found only 4 false hits, suggesting a precision of 96%. Most cases (8,031) are with *waarvan* as relative clause head, 1,490 have *van wie* as relative clause head. The complement clause is usually a regular verbal complement, but sometimes (1,488 cases) functions as subject. The complementizer is almost always *dat* (9,062 cases), but examples with complementizer *of* and *alsof* occur as well (459 cases).

The distribution of matrix verbs and matching resumptive pronouns is given in table 4. The two most frequent verbs are *denken*, which is most frequent for true, and *weten*, for which it is usually claimed that it cannot occur in long-distance dependencies. The data confirms the observation in Hoeksema and Schippers (2012) that this construction is not subject to island constraints: there is a wide variety in matrix verbs, most of which are not known to be 'bridge verbs', in 459 cases the resumptive pronoun is in a complement clause headed by *(als)of,* and in 1,488 the resumptive pronoun is in a subject clause. The latter are mostly cases involving the copula *zijn:*

(13) Soorten **waarvan** het onduidelijk is of **ze** in Nederland voorkomen
species of-which it unclear is whether they in the Netherlands occur
'species of which it is unclear whether they occur in the Netherlands'

7 R-Pronominal Parasitic gaps

In the previous sections we have been concerned with searching for true LDDs in an annotated corpus, and searching for a popular strategy for avoiding such dependencies. In this section we add some observations on a closely related construction that seems to be extremely scarce in actual data as well.

In Dutch, non-local dependencies between a fronted WH-element and a position governed by a preposition are in general not allowed. So-called *'R-pronouns'* (following the discussion in van Riemsdijk [1978]) are an exception to this rule. They can be used both to form WHquestions, as in (14-b), as well as discontinuous constituents where the r-pronoun precedes but is non-adjacent to its governing preposition (14-d).

(14) a. *Wat ben je voor verzekerd?
What are you for insured
b. Waar ben je voor verzekerd?
What[+R] are you for insured
'What are you insured for¿
c. *Je bent het niet voor verzekerd
You are it not for insured
d. Je bent er niet voor verzekerd
You are it[+R] not for insured
'You are not insured for it'

A recent paper (Everaert et al. 2015) arguing for structure being more prominent than word order in syntax uses this construction to produce Dutch example sentences like (15-b).

(15) a. Ik ben speciaal voor het klimaat naar de Provence toe gereden
I am especially for the climate to the Provence driven
'I drove to Provence especially for the climate'
b. Ik ben **er** speciaal **voor naar toe** vertrokken
I am it especially for to to driven
'I drove there especially for it'

Compared to (15-a), which does contain two full PPs, the R-pronoun *er* in (15-b) seems to be dependent on a gap in two PPs. Everaert et al. (2015) draw a parallel between cases such as this and parasitic gap constructions (Engdahl 1983). The examples were discussed in a blog[6] that sparked a lively discussion, including a response by one of the authors of the original paper.[7]

6 http://nederl.blogspot.nl/2015/11/ik-ben-er-speciaal-voor-naartoe-gereden.html
7 http://nederl.blogspot.nl/2015/11/recursie-en-evolutie-van-taal.html

While this construction does not involve a true long-distance dependency, we include it in our discussion as it does involve a rare construction involving non-local dependencies.

Huijbrechts (p.c., Huijbrechts [2016]) presents additional examples such as (16).

(16) a. **Waar** rekent hij **op** om **naar toe** te gaan?
Where counts he on PRT to to to go
'Where does he count on to go to?'
b. **Waar** ga je **van** uit dat zij **op** zal letten?
Where go you from out that she on will note
'What do you suppose she will pay attention to?'

These constructions are a slight variation of the R-pronominal parasitic gap constructions in (15-b), in that they involve a gap in a PP in a complement clause, and a suppressed R-pronoun in the main clause. Note that normally, PPs containing a complement clause are obligatorily introduced by the expletive R-pronoun er:

(17) a. Hij rekent er op om naar Amsterdam toe te gaan
He counts there on CMP to Amsterdam to to go
'He counts on going to Amsterdam'
b. Je gaat er van uit dat zij op schrijffouten zal letten?
You go there of out that she on spelling-errors will notice
'You are counting on her to pay attention to spelling errors'

One of the questions is to what extent such phenomena occur in spontaneous data. If not, or scarcely, they constitute evidence for a 'Poverty of the Stimulus' argument: apparently, language users are able to produce and understand parasitic gap constructions without necessarily having been exposed to such sentences in the past.

One problem with this argument is that it is very hard to check for the occurrence of configurations such as (15-b) and (16) in corpora. The Alpino parser, while based on a linguistically sophisticated hand-written grammar, does not cover parasitic gap constructions. As a consequence, these will not be analyzed as such in corpora that are analyzed automatically by Alpino. Given a sufficiently large corpus, one might search for sentences containing the trigram *voor naar toe* and check these manually. The NL-COW corpus (text from Dutch language websites, 259M sentences)[8] contains 19 occurrences of the string *voor naartoe*,[9] of which at least a few cases are similar to the example presented by Everaert et al. (2015):

8 http://corporafromtheweb.org
9 We opted for searching for the more common spelling *naartoe over naar toe.*

(18) a. … ik zou **er** niet speciaal voor naartoe gaan
… I would there not especially for towards go
… 'I would not especially go there for it'
b. **Er** speciaal **voor naartoe** rijden hoefde niet
There especially for towards drive needed not
'It was not necessary to drive there for it especially'

However, this kind of search is very limited, as (1) it presupposes that the two prepositions are adjacent, which need not be the case in parasitic gap constructions in general, and (2) it fails to check for cases involving other prepositions.

Another possibility is spotting such constructions 'in the wild'. For instance, after becoming aware of examples such as (16), we noticed the following quote:[10]

(19) Daar heb je dan geen tijd voor om naar te kijken
there have you than no time for cmp to to watch
At that moment, you do not have time to look at that

This suggests that maybe constructions like these have simply gone unnoticed by linguists.

A more effective strategy involves searching for potential parasitic gaps in Lassy Large. As Alpino does not take parasitic gaps into account, we will have to formulate a query that only approximates the relevant syntactic configuration, and check results manually. We used the following query:

(20)
```
//node[node[@rel="rhd" and @lemma="waar"] and
    descendant::node[node[@cat="pp"]/node[@index and not(@pos or @cat)]
                     descendant::node[@rel="vc" and
                                      (@cat="ti" or @cat="oti" or @cat="cp")
                                          ]
                    ]
      ]
```

Here, we search for sentences containing a relative clause headed by *waar*, and containing a PP containing a gap, and a complement clause. Such sentences might, but are not guaranteed to, contain the relevant structure.

The query gives rise to 564 hits on Lassy Large, of which 16 cases appear to be instances of the phenomenon we are interested in. Two examples are given below:

10 Interview with cyclist Matteo Trentin (translated into Dutch) by Nando Broers in *De Muur*, 2016/2.

(21) a. Het soort waar iedere vrouw van zou moeten dromen
The kind of-which every woman of should must dream
om te trouwen
COMP to marry
'the kind which every woman should dream of to marry with'

b. Dit zijn de genen waar men voor heeft gekozen om onderzoek
These are the genes which one for has chosen COMP research
naar te doen
into to do
'These are the genes for which one has chosen to do research on'

The results of the query are very noisy. Although it may be possible to modify the query to achieve slightly better precision, we do believe that these constructions are very hard to detect in the output of the current Alpino grammar. In terms of frequency, examples like these do seem almost as frequent as long-distance dependencies in relative clauses containing a gap in a tensed subordinate clause or in a complement clause introduced by *om*.

8 Conclusions

In this paper, we have searched for true long-distance dependencies in an annotated corpus. True LDDs in relatives and WH-questions containing a subordinate clause (either tensed or introduced by the complementizer *om* or containing an 'extraposed' infinitival complement) are all covered by the Alpino parser, and thus can be searched for directly. Manual inspection of the results was necessary as the precision of the parser on these constructions is not very high. The results show that true LDDs are quite infrequent in the corpus but do seem to provide support for claims that there are collocational effects in this construction.

Two related constructions, resumptive prolepsis and R-pronominal parasitic gaps, are outside the scope of the grammar. For the resumptive pronoun construction, an approximate query turned out to be quite accurate, and gave rise to a high number of results. The distribution of matrix verbs in this construction supports the findings of Hoeksema and Schippers (2012). For R-pronominal parasitic gaps, it is much harder to come up with a good approximate query. However, after manual filtering we did find a number of positive examples. In this case, the main advantage of using a syntactically annotated corpus is that it makes it possible to search somewhat efficiently for this phenomenon in the first place.

The Lassy Large corpus seems sufficiently large and heterogeneous to support research on long-distance dependencies, and the automatic syntactic annotation,

while far from perfect, does help to zoom in on the interesting cases quickly. Several questions remain for further research, such as estimating the recall of the automatic parser, and collecting statistics for other longdistance dependency constructions, such as comparatives.

References

Augustinus, Liesbeth, Vandeghinste, Vincent, Schuurman, Ineke and van Eynde, Frank. 2017. GrETEL: A tool for example-based treebank mining. In Jan Odijk and Arjan van Hessen (eds.), *Clarin in the Low Countries*, 269–280. London: Ubiquity Press.

Bennis, Hans. 2000. Adjectives and argument structure. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 27–68.

Bouma, Gosse. 2017. Om-omission. In Martijn Wieling, Martin Kroon, Gosse Bouma and Gertjan van Noord (eds.), *From Semantics to dialectometry: Festschrift for John Nerbonne*, 65–74. London: College Publications.

Bouma, Gosse, Malouf, Rob and Sag, Ivan. 2001. Satisfying Constraints on Adjunction and Extraction. *Natural Language and Linguistic Theory* 19, 1–65.

Bouma, Gosse and van Noord, Gertjan. 2017. Increasing return on annotation investment: the automatic construction of a Universal Dependency treebank for Dutch. In Joachim Nivre and Marie-Catherine de Marneffe (eds.), *NoDaLiDa workshop on Universal Dependencies*, Gothenburg.

Broekhuis, Hans, Den Besten, Hans, Hoekstra, Kees and Rutten, Jean. 1995. Infinitival complementation in Dutch: On remnant extraposition. *The Linguistic Review* 12(93-122).

Candito, Marie and Seddah, Djam'e. 2012. Effectively long-distance dependencies in French: annotation and parsing evaluation. In Iris Hendrickx, Sandra Kübler and Kiril Simov (eds.), *TLT 11 – The 11th International Workshop on Treebanks and Linguistic Theories*.

Chomsky, Noam. 1977. On wh-movement. In Akmajian Adrian Culicover Peter, Wasow Thomas (ed.), *Formal Syntax*. New York: Academic Press.

Coppen, Peter-Arno. 2013. De zin die wij merken dat ook voor linguïstische problemen zorgt. *Nederlandse Taalkunde* 18(2), 193–203.

Cremers, Crit. 1983. On two types of infinitival complementation. In Frank Heny (ed.), *Linguistic categories: auxiliaries and related puzzles*. 169–221, Dordrecht: Springer.

Engdahl, Elisabet. 1983. Parasitic gaps. *Linguistics and philosophy* 6(1), 5–34.

Everaert, Martin, Huybregts, Marinus, Chomsky, Noam, Berwick, Robert and Bolhuis, Johan. 2015. Structures, not Strings: Linguistics as part of the Cognitive Sciences. *Trends in Cognitive Sciences* 19, 729–743.

Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey and Sag, Ivan. 1985. *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Hoeksema, Jack and Schippers, Ankelien. 2012. Diachronic changes in long-distance dependencies. *Historical Linguistics 2009: Selected Papers from the 19th International Conference on Historical Linguistics, Nijmegen, 10–14 August 2009,* 155–170.

Hofmeister, Philip and Sag, Ivan A. 2010. Cognitive constraints and island effects. *Language* 86 (2), 366–415.

Huijbrechts, Riny. 2016. Binding Unleashed. Ms. Utrecht University.

Kaplan, Ronald M. and Zaenen, Annie. 1989. Long-distance Dependencies, Constituent Structure and Functional Uncertainty. In Mark R. Baltin and Anthony S. Kroch (eds.), *Alternative Conceptions of Phrase Structure*, University of Chicago Press.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

Morrill, Glyn. 1995. Discontinuity in categorial grammar. *Linguistics and Philosophy* 18(2), 175–219.

Odijk, Jan. 2015. Linguistic Research with PaQu. *Computational Linguistics in The Netherlands journal* 5, 3–14.

Oostdijk, Nelleke, Reynaert, Martin, Hoste, Véronique and Schuurman, Ineke. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk (eds.), *Essential speech and language technology for Dutch*, 219– 247. Heidelberg: Springer.

Rimell, Laura, Clark, Stephen and Steedman, Mark. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Volume 2, 813–821, Association for Computational Linguistics.

Ross, J.R. 1967. *Constraints on variables in syntax*. Ph. D.thesis, Massachusetts Institute for Technology.

Salzmann, Martin. 2006. *Resumptive Prolepsis: A Study in Indirect A' Dependencies*. Ph. D.thesis, Leiden University, Leiden.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic inquiry* 31(4), 649–689.

van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister and Patrick Watrin (eds.), *TALNo6. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42.

van Noord, Gertjan, Bouma, Gosse, van Eynde, Frank, de Kok, Daniel, van der Linde, Jelmer, Schuurman, Ineke, Sang, Erik Tjong Kim and Vandeghinste, Vincent. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In

Peter Spyns and Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch: the STEVIN Programme*, 147–164, Springer.

van Riemsdijk, Henk. 1978. *A Case Study in Syntactic Markedness: The binding nature of prepositional phrases*. Foris Publications, Dordrecht.

Verhagen, Arie. 2006. On subjectivity and 'long distance Wh-movement'. In Angeliki Athanasiadou, Costas Canakis and Bert Cornillie (eds.), *Subjectification: Various Paths to Subjectivity*, 323–346, Berlin: Mouton de Gruyter.