Parsed Corpora for Linguistics

Gertjan van Noord
University of Groningen
G.J.M.van.noord@rug.nl

Gosse Bouma
University of Groningen
G.Bouma@rug.nl

Abstract

Knowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically. We argue that such parsed corpora are an interesting new resource for linguists. The argument is illustrated by means of a number of recent results which were established with the help of parsed corpora.

1 Introduction

Once upon a time, knowledge-based parsers were slow, inaccurate and fragile. This is no longer true. In the last decade, enormous improvements have been achieved in this area. Parsers based on constraint-based formalisms such as HPSG, LFG, and CCG are now fast enough for many applications; they are robust; and they perform much more accurately than previously by incorporating, typically, a statistical disambiguation component. As a consequence, such parsers now obtain competitive, if not superior, performance. Zaenen (2004), for instance, points out that the (LFG-based) XLE parser is fast, has a statistical disambiguation component, and is robust, and thus allows full parsing to be incorporated in many applications. Clark and Curran (2007) show that both accurate and highly efficient parsing is possible using a CCG.

As a consequence of this development, massive amounts of parsed sentences now become available. Such large collections of syntactically annotated but not manually verified syntactic analyses are a very useful resource for many purposes. In this position paper we focus on one purpose: linguistic analysis. Our claim is, that very large *parsed* corpora are an important resource for linguists. Such very large parsed corpora can be used to search systematically for specific infrequent syntactic configurations of interest, and also to obtain quantitative data about specific syntactic configurations. Although parsed corpora obviously contain a certain amount of noise, for many applications the abundant size of these corpora compensates for this.

In this paper, we illustrate our position by a numer of recent linguistic studies in which very large corpora of Dutch have been employed, which were syntactically annotated by the freely available Alpino parser (Bouma et al., 2001; van Noord, 2006).

The Alpino system incorporates a linguistically motivated, wide-coverage grammar for Dutch in the tradition of HPSG. It consists of over 800 grammar rules and a large lexicon of over 300,000 lexemes (including very many person names, geographical names, and organization names) and various rules to recognize special constructs such as named entities, temporal expressions, etc. Since we use Alpino to parse large amounts of data, it is crucial that the parser is capable to treat sentences with unknown words. A large set of heuristics have been implemented carefully to deal with unknown words and word sequences.

Based on the categories assigned to words, and the set of grammar rules compiled from the HPSG grammar, a left-corner parser finds the set of all parses, and stores this set compactly in a packed parse forest. All parses are rooted by an instance of the top category, which is a category that generalizes over all maximal projections (S, NP, VP, ADVP, AP, PP and some others). If there is no parse covering the complete input, the parser finds all parses for each substring. In such cases, the robustness component will then select the best sequence of non-overlapping parses (i.e., maximal projections) from this set.

In order to select the best parse from the parse forest, a best-first search algorithm is applied. The algorithm consults a Maximum Entropy disambiguation model to judge the quality of (partial)

Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, pages 33–39,
Athens, Greece, 30 March, 2009.

parses. The disambiguation model is trained on the manually verified Alpino treebank (about 7100 sentences from newspaper texts).

Although Alpino is not a dependency grammar in the traditional sense, dependency structures are generated by the lexicon and grammar rules as the value of a dedicated feature. The dependency structures are based on CGN (Corpus Gesproken Nederlands, Corpus of Spoken Dutch) (Hoekstra et al., 2003), D-Coi and LASSY (van Noord et al., 2006).

Dependency structures are stored in XML. Advantages of the use of XML include the availability of general purpose search and visualization software. For instance, we exploit XPATH (standard XML query language) to search in large sets of dependency structures, and Xquery to extract information from such large sets of dependency structures (Bouma and Kloosterman, 2002; Bouma and Kloosterman, 2007).

2 Extraposition of comparative objects out of topic

The first illustration of our thesis that parsed corpora provide an interesting new resource for linguists, constitutes more of an anecdote than a systematic study. We include the example, presented earlier in van Noord (2009), because it is fairly easy to explain, and because it was how we became aware ourselves of the potential of parsed corpora for the purpose of linguistics.

In van der Beek et al. (2002), the grammar underlying the Alpino parser is presented in some detail. As an example of how the various specific rules of the grammar interact with the more general principles, the analysis of comparatives and the interaction with generic principles for (rightward) extraposition is illustrated. In short, comparatives such as comparative adjectives and the adverb `anders` as in the following example (1) license corresponding comparative phrases (such as phrases headed by `dan` (than)) by means of a feature which percolates according to the extraposition principle. The analysis is illustrated in figure 1.

(1) . . . niks anders doen dan almaar
. . . nothing else do than continuously
ruw materiaal verzamelen
raw material collect
do nothing else but collect raw material (cdbl-7)

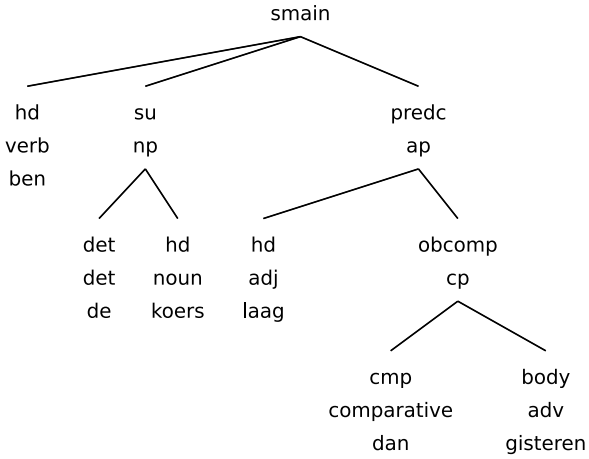

Figure 2: Dependency structure for *Lager was de koers dan gisteren*

An anonymous reviewer criticized the analysis, because the extraposition principle would also allow the rightward extraction of comparative phrases licensed by comparatives in topic position. The extraposition principle would have to allow for this in the light of examples such as

(2) De vraag is gerechtvaardigd waarom de
The question is justified why the
regering niets doet
government nothing does
The question is justified why the goverment does not act

However, the reviewer claimed that comparative phrases cannot be extraposed out of topic, as examples such as the following indicate:

(3) *Lager was de koers dan gisteren
Lower was the rate than yesterday
The rate never was lower than yesterday

Since the Alpino grammar allows such cases, it is possible to investigate if genuine examples of this type occur in parsed corpora. In order to understand how we can specify a search query for such cases, it is instructive to consider the dependency structure assigned to such examples in figure 2. As can be observed in the dependency graph, the left-right order of nodes does not represent the left-right ordering in the sentence. The word-order of words and phrases is indicated with XML attributes *begin* and *end* (not shown in figure 2) which indicate for each node the begin and end position in the sentence respectively.

The following XPATH query enumerates all ex-

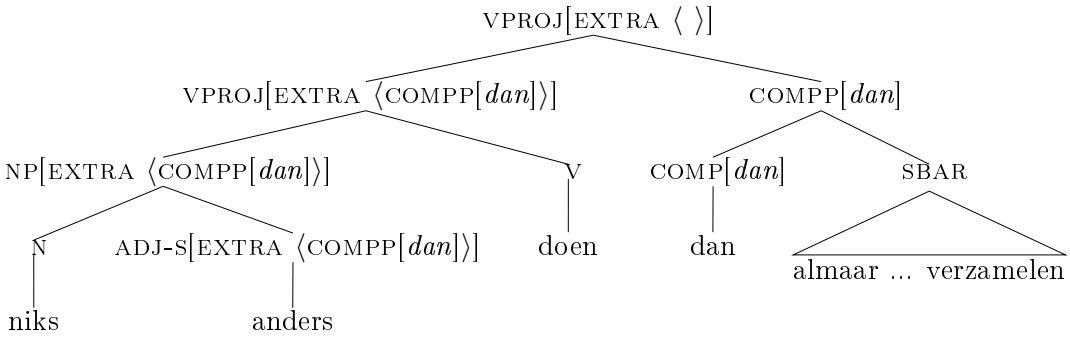

Figure 1: Derivation of extraposed comparative object

amples of extraposition of comparative phrases out of topic. We can then inspect the resulting list to check whether the examples are genuine.

```
//node[
    @cat="smain"
    and
    node[
      node[@rel="obcomp"]/@end
      >
      ../node[@rel="hd"]/@begin
        ]/@begin = @begin
       ]
```

The query can be read as: find root sentences in which there is a daughter node, which itself has a daughter node with relation label `obcomp` (the label used for comparative complements). The daughter node should begin at the same position as the root sentence. Finally, the end position of the `obcomp` node must be larger than the end position of the head of the root sentence (i.e. the finite verb).

In addition to many mis-parsed sentences, we found quite a few genuine cases. A mis-parse can for instance occur if a sentence contains two potential licensers for the comparative phrase, as in the following example in which *verder* can be wrongly analysed as a comparative adjective.

(4) Verder wil ik dat mijn backhand even
Further want I that my backhand just-as
goed wordt als mijn forehand
good becomes as my forehand
Furthermore, I want my backhand to become as good as my forehand

More interestingly for the present discussion are the examples which were parsed correctly. Not only do we find such examples, but informants agree that nothing is wrong with such cases. Some examples are listed in figure 3. It is striking that many examples involve the comparative adjectives *liever* and *eerder*. Also, the list involves examples where adverbials such as *zo, zozeer, zoveel* are related with an extraposed subordinate sentence headed by *dat* which according to the annotation guidelines are also treated as comparative complements.

The examples show that at least in some cases, the possibility of extraposition of comparative complements out of topic must be allowed; we hypothesize that the acceptability of such cases is not a binary decision, but rather a preference which depends on the choice of comparative on the one hand, and the heaviness of the comparative complement on the other hand.

For the purpose of this paper, we hope to have illustrated how parsed corpora can be helpful to find new empirical evidence for fairly complicated and suble linguistic issues. Note that for a construction of this type, manually verified treebanks are much too small. We estimated that it takes about 5 million words to find a single, good, example. It appears unrealistic to assume that treebanks of the required order of magnitude of tens of millions of words will become available soon.

3 Frequency versus Complexity

Our second illustration is of a different nature, and taken from a study related to agrammatic Broca's aphasia.

In Bastiaanse et al. (to appear), potential causes are discussed of the problems that patients suffering from agrammatic Broca's aphasia encounter. The *Derived Order Problem Hypothesis* (Bastiaanse and van Zonneveld, 2005) assumes that the linguistic representations of agrammatic patients are intact, but due to processing disorders, some representations are harder to retrieve than oth-

(5) Liever betaalden werkgevers een (hoge) verzekeringspremie , dan opgescheept te zitten met niet
Rather paid employers a (high) insurance-fee , than left to be with not
volwaardig functionerende medewerkers
fully functioning employees
Rather, employers pay a high insurance fee, than be left with not fully functioning employees (Algemeen Dagblad, January 15, 1999)

(6) Beter is het te zorgen dat ziekenhuizen hun verplichtingen volgens de huidige BOPZ gaan
Better is it to ensure that hospitals their obligations according-to the current BOPZ start
nakomen , dan de rechten van patiï¿½ten nog verder aan te tasten
meet , than the rights of patients yet further PART to violate
It is better to ensure that hospitals start to meet their obligations according to the current BOZP, than to violate rights of patients even further (Algemeen Dagblad, August 18, 2001)

(7) Dus wat anders konden de LPF'ers de afgelopen week dan zich stil houden ?
So what else could the LPF-representatives the last week than self quiet keep ?
What else could the LPF-representatives do last week , than keep quiet? (Volkskrant June 1, 2002)

(8) Sneller kennen ze hun tafels van vermenigvuldiging dan de handelingen van de groet
Faster know they their tables of multiplication than the acts of the greeting
They know the tables of multiplication faster than the acts of greeting (De Morgen March 27, 2006)

Figure 3: Some genuine examples of extraposition of comparative objects from topic. The examples are identified automatically using an XPATH query applied to a large parsed corpus.

ers, due to differences in linguistic complexity. This hypothesis thus assumes that agrammatic patients have difficulty with constructions of higher linguistic complexity. An alternative hypothesis states, that agrammatic patients have more difficulty with linguistic constructions of lower frequency.

In order to compare the two hypotheses, Bastiaanse *et al.* perform three corpus studies. In three earlier experimental studies it was found that agrammatic patients have more difficulty with (a) finite verbs in verb-second position versus finite verbs in verb-final position; (b) scrambled direct objects versus non-scrambled direct objects; and (c) transitive verbs used as unaccusative versus transitive verbs used as transitive.

The three pairs of constructions are illustrated as follows.

(9) a. de jongen die een boek leest
the boy who a book reads
the boy who reads a book
b. de jongen leest een boek
the boy reads a book
the boy reads a book

(10) a. dit is de jongen die vandaag het boek
this is the boy who today the book
leest
reads
this is the boy who reads the book today
b. dit is de jongen die het boek vandaag
this is the boy who the book today
leest
reads
this is the boy who reads the book today

(11) a. de jongen breekt het glas
the boy breaks the glass
the boy breaks the glass
b. het glas breekt
the glass breaks
the glass breaks

In each of the three cases, corpus data is used to estimate the frequency of both syntactic configurations. Two corpora were used: the manually verified syntactically annotated CGN corpus (spoken language, approx. 1M words), and the the automatically parsed TwNC corpus (Ordelman et al., 2007) (the newspapers up to 2001, a parsed corpus of 300 million words). For the first two experiments, manual inspection revealed that the parsed corpus material was of high enough quality to be used directly. Furthermore, the relevant constructions are highly frequent, and thus even relatively small corpora (such as the syntactically an-

notated part of CGN) provide sufficient data. For the third experiment (unaccusative versus transitive usage of verbs), an additional layer of manual verification was used, and furthermore, as the subcategorization frequencies of individual verbs are estimated, the full TwNC was searched in order to obtain reasonably reliable estimates.

The outcome of the three experiments was the same in each case: frequency information cannot explain the difficulty encountered by agrammatic patients. Verb-second is more frequent than verb-final word order for lexical verbs and transitive lexical verbs (the verbs used in the experiments were all transitive). Finite verbs occur slightly more often in verb-second position than in verb-final position, but the difference is quite small. Scrambled word order is more frequent than the basic word order. The difference between the two corpora (CGN and TwNC) is quite small in both cases. Figure 4 gives an overview of the number of occurrences of the transitive and unaccusative use of the verbs used in the experiments in the full TwNC. The data suggest that the relative frequency of unaccusative depends strongly on the verb, but that it is not in general the case that the unaccusative use is less frequent than the transitive use.

The three 'difficult' constructions used in the experiments with aphasia patients are by no means infrequent in Dutch. The authors conclude that the hypothesis that processing difficulties are correlated with higher linguistic complexity cannot be falsified by an appeal to frequency.

What is interesting for the purposes of the current paper, is that parsed corpora are used to estimate frequencies of syntactic constructions, and that these are used to support claims about the role of linguistic complexity in processing difficulties of aphasia patients. Also note that figure 4 shows that even in a large (300M word) corpus, the number of occurrences of a specific verb used with a specific valency frame can be quite small. Thus, it is unlikely that reliable frequency estimates can be obtained for these cases from manually verified treebanks.

Roland et al. (2007) report on closely related work for English. In particular, they give frequency counts for a range of syntactic constructions in English, and subcategorization frequencies for specific verbs. They demonstrate that these frequencies are highly dependent on corpus and genre in a number of cases. They use their data to verify claims in the psycholinguistic literature about the processing of subject vs. object clefts, relative clauses and sentential complements.

4 The distribution of *zelf* and *zichzelf*

As a further example of the use of parsed corpora to further linguistic insights, we consider a recent study (Bouma and Spenader, 2009) of the distribution of weak and strong reflexive objects in Dutch.

If a verb is used reflexively in Dutch, two forms of the reflexive pronoun are available. This is illustrated for the third person form in the examples below.

(12) Brouwers schaamt **zich/*****zichzelf** voor zijn schrijverschap.
Brouwers shames *self1/self2* for his writing
Brouwers is ashamed of his writing

(13) Duitsland volgt ***zich/zichzelf** niet op als Europees kampioen.
Germany follows *self1/self2* not PART as European Champion
Germany does not succeed itself as European champion

(14) Wie **zich/zichzelf** niet juist introduceert, valt af.
Who self1/self2 not properly introduces, is out
Everyone who does not introduce himself properly, is out.

The choice between *zich* and *zichzelf* depends on the verb. Generally three groups of verbs are distinguished. Inherent reflexives are claimed to never occur with a non-reflexive argument, and as a reflexive argument are claimed to use *zich* exclusively, (12). Non-reflexive verbs seldom, if ever occur with a reflexive argument. If they do however, they can only take *zichzelf* as a reflexive argument (13). Accidental reflexives can be used with both *zich* and *zichzelf*, (14). Accidental reflexive verbs vary widely as to the frequency with which they occur with both arguments. Bouma and Spenader (2009) set out to explain this distribution.

The influential theory of Reinhart and Reuland (1993) explains the distribution as the surface realization of two different ways of reflexive coding. An accidental reflexive that can be realized with

verb		unacc		trans	
		#	%	#	%
luiden	*to ring/sound*	269	26.6	743	73.4
scheuren	*to rip*	332	28.8	819	71.2
breken	*to break*	1969	31.2	4341	68.8
verbrand	*to burn*	479	43.5	623	56.5
oplossen	*to (dis)solve*	296	59.2	204	40.8
draaien	*to turn*	2709	59.4	1852	40.6
smelten	*to melt*	723	71.4	290	28.6
rollen	*to roll*	3500	93.5	244	6.5
verdrinken	*to drown*	1397	94.6	80	5.4
stuiteren	*to bounce*	334	97.9	7	2.1

Figure 4: Estimated number of occurrences in TwNC of unaccusative and transitive uses of Dutch verbs which may undergo the causative alternation

both *zich* and *zichzelf* is actually ambiguous between an inherent reflexive and an accidental reflexive (which always is realized with *zichzelf*). An alternative approach is that of Haspelmath (2004), Smits et al. (2007), and Hendriks et al. (2008), who have claimed that the distribution of weak vs. strong reflexive object pronouns correlates with the proportion of events described by the verb that are self-directed vs. other-directed.

In the course of this investigation, a first interesting observation is, that many inherently reflexive verbs, which are claimed not to occur with *zichzelf*, actually often do combine with this pronoun. Here are a number of examples (simplified for expository purposes):

(15) Nederland moet stoppen zichzelf op de
Netherlands must stop self2 on the
borst te slaan
chest to beat
The Netherlands must stop beating itself on the chest

(16) Hunze wil zichzelf niet al te zeer op
Hunze want self2 not all too much on
de borst kloppen
the chest knock
Hunze doesn't want to knock itself on the chest too much

(17) Ze verloren zichzelf soms in het
They lost self2 sometimes in tactical
gegoochel met allerlei tactische varianten
variants
They sometimes lost themselves in tactical variants

With regards to the main hypothesis of their study, (Bouma and Spenader, 2009) use linear regression to determine the correlation between reflexive use of a (non-inherently reflexive) verb and the relative preference for a weak or strong reflexive pronoun. Frequency counts are collected from the parsed TwNC corpus (almost 500 million words). They limit the analysis to verbs that occur at least 10 times with a reflexive meaning and at least 50 times in total, distinguishing uses by subcategorization frames. The statistical analysis shows a significant correlation, which accounts for 30% of the variance of the ratio of nonreflexive over reflexive uses.

5 Conclusion

Knowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically. We argued that such parsed corpora are an interesting new resource for linguists. The argument is illustrated by means of a number of recent results which were established with the help of huge parsed corpora.

Huge parsed corpora are especially crucial (1) to obtain evidence concerning infrequent syntactic configurations, and (2) to obtain more reliable quantitative data about particular syntactic configurations.

Acknowledgments

This research was carried out in part in the context of the STEVIN programme which is funded by the Dutch and Flemish governments

(http://taalunieversum.org/taal/technologie/stevin/).

References

Roelien Bastiaanse and Ron van Zonneveld. 2005. Sentence production with verbs of alternating transitivity in agrammatic Broca's aphasia. *Journal of Neurolinguistics*, 18(1):57–66, January.

Roelien Bastiaanse, Gosse Bouma, and Wendy Post. to appear. Frequency and linguistic complexity in agrammatic speech production. *Brain and Language*.

Gosse Bouma and Geert Kloosterman. 2002. Querying dependency treebanks in XML. In *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, pages 1686–1691, Gran Canaria, Spain.

Gosse Bouma and Geert Kloosterman. 2007. Mining syntactically annotated corpora using XQuery. In *Proceedings of the Linguistic Annotation Workshop*, Prague, June. ACL.

Gosse Bouma and Jennifer Spenader. 2009. The distribution of weak and strong object reflexives in Dutch. In Frank van Eynde, Anette Frank, Koenraad De Smedt, and Gertjan van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, number 12 in LOT Occasional Series, pages 103–114, Utrecht, The Netherlands. Netherlands Graduate School of Linguistics.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Wide coverage computational analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Computational Linguistics in the Netherlands 2000*.

S. Clark and J.R. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33(4):493–552.

Martin Haspelmath. 2004. A frequentist explanation of some universals of reflexive marking. Draft of a paper presented at the Workshop on Reciprocals and Reflexives, Berlin.

Petra Hendriks, Jennifer Spenader, and Erik-Jan Smits. 2008. Frequency-based constraints on reflexive forms in Dutch. In *Proceedings of the 5th International Workshop on Constraints and Language Processing*, pages 33–47, Roskilde, Denmark.

Heleen Hoekstra, Michael Moortgat, Bram Renmans, Machteld Schouppe, Ineke Schuurman, and Ton van der Wouden, 2003. *CGN Syntactische Annotatie*, December.

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. 2007. TwNC: a multifaceted Dutch news corpus. *ELRA Newsletter*, 12(3/4):4–7.

Tanya Reinhart and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry*, 24:656–720.

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

Erik-Jan Smits, Petra Hendriks, and Jennifer Spenader. 2007. Using very large parsed corpora and judgement data to classify verb reflexivity. In Antonio Branco, editor, *Anaphora: Analysis, Algorithms and Applications*, pages 77–93, Berlin. Springer.

Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374.

Gertjan van Noord, Ineke Schuurman, and Vincent Vandeghinste. 2006. Syntactic annotation of large corpora in STEVIN. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

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

Gertjan van Noord. 2009. Huge parsed corpora in Lassy. In Frank van Eynde, Anette Frank, Koenraad De Smedt, and Gertjan van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, number 12 in LOT Occasional Series, pages 115–126, Utrecht, The Netherlands. Netherlands Graduate School of Linguistics.

Annie Zaenen. 2004. but full parsing is impossible. *ELSNEWS*, 13(2):9–10.