

Huge Parsed Corpora in LASSY

Gertjan van Noord
University of Groningen
E-mail: G. J. M. van.Noord@rug.nl

Abstract

One of the goals of the LASSY STEVIN project (Large Scale Syntactic Annotation of written Dutch) is a syntactically annotated (manually verified) corpus of 1 million words. In addition, the full STEVIN reference corpus of 500 million words will be syntactically annotated automatically. In this paper, the potential of such huge treebanks for applications in corpus linguistics, natural language processing and information extraction is illustrated.

1 Introduction

The construction of a 500 million word reference corpus of written Dutch has been identified as one of the priorities in the STEVIN programme. Parts of this corpus will be enriched with verified linguistic annotations. The goal of the LASSY STEVIN project is to extend the amount of syntactically annotated material. At the end of LASSY there will be a manually verified treebank consisting of 1 million words. Perhaps more interestingly, the full 500 million word reference corpus will be syntactically annotated automatically, and the resulting huge treebank will be made available as well. The outcome of LASSY therefore contains two types of treebank: a relatively small treebank of high quality on the one hand, and a huge treebank of lower quality on the other hand.

The availability of small quantity, high quality treebanks for research and development in natural language processing is crucial, in particular for training statistical syntactic analysers or statistical components of syntactic analysers of a more hybrid nature. In addition, small quantity, high quality treebanks are important for evaluation purposes for any kind of automatic syntactic analysers.

It is less obvious whether large quantity, lower quality treebanks are a useful resource. In this paper, we illustrate the expected quality of the automatic annotations, and we provide a number of example studies which illustrate the promise of large quantity, lower quality treebanks in areas such as information extraction,

ontology construction, lexical acquisition, corpus linguistics, and natural language processing itself.

2 Automatically Parsed Treebanks

In LASSY, we use the freely available Alpino parser for Dutch [13]. Alpino is computational analyzer of Dutch which aims at full accurate parsing of unrestricted text, and which incorporates both knowledge-based techniques, such as a HPSG-grammar and -lexicon which are both organized as inheritance networks, as well as corpus-based techniques, for instance for training its POS-tagger and its Maximum Entropy disambiguation component.

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 *dt*. The dependency structures are based on CGN (Corpus Gesproken Nederlands, Corpus of Spoken Dutch) [7], D-Coi and LASSY [15].

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.

The output of the parser is evaluated by comparing the generated set of named dependencies for a corpus sentence to the set of named dependencies extracted from the manually annotated treebank. Comparing these sets, we count the number of dependencies that are identical in the generated parse and the stored structure. The concept accuracy (CA) measure indicates the proportion of correct named dependencies.

In order to judge the expected quality of automatically parsed corpora, we list the concept accuracy figures for the manually verified sub-corpora developed in STEVIN D-Coi. For these sub-corpora, we compare the dependency structures produced by Alpino with the manually verified dependency structures. The results are aggregated over the various sub-corpora, because it can be expected that the accuracy of syntactic dependencies differ with respect to domain. As can be observed in the table, most sub-corpora can be parsed with an accuracy well over 80%. Newspaper texts are particularly easy for Alpino, because Alpino is trained on the Alpino treebank, which also consists of newspaper text. The *books* sub-corpus was relatively easy, because the corpus sentences all came from a single children's book. Legal texts constitute the hardest sub-domain in this collection.

material	number of words
XMLWiki Wikipedia 2008	110M
Europarl version 3	38M
TwNC newspaper material	531M
Mediargus newspaper material	1397M
European Medicines Agency	14M

Table 1: Some of the corpora which have been automatically parsed in the context of the Lassy project.

sub-corpus	# sentences	CA%	sub-corpus	# sentences	CA%
newsletters	90	83.17	web sites	2833	86.56
wikipedia	240	88.38	press releases	591	84.46
books	276	92.86	brochures	1796	85.84
flyers	306	87.93	manuals	397	80.26
legal texts	279	76.96	newspaper	2267	91.03
policy docs	1266	84.73	proceedings	339	88.08
reports	1115	88.60	<i>total</i>	12637	86.52

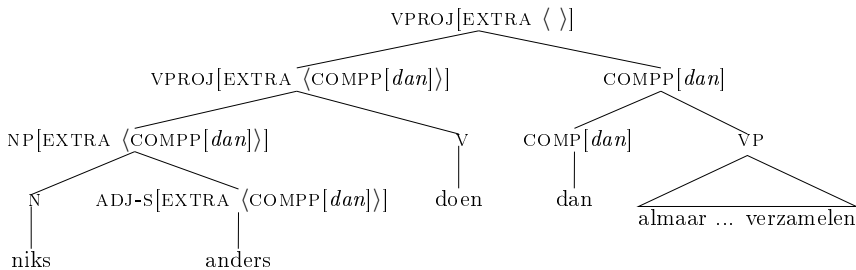
In Lassy, the full STEVIN reference corpus of 500 million words will be automatically annotated. This corpus is not yet available. In the experiments reported in this paper, several other large automatically annotated corpora have been used. Table 1 lists some of the automatically annotated corpora which are currently available.

3 Corpus Linguistics

Huge parsed corpora are useful for research in corpus linguistics. In this section, we will illustrate this by means of an anecdotal example. However, there are also examples of more systematic corpus linguistics research, exploiting huge corpora parsed with Alpino (one such example is presented in Bouma & Spenader, 2009 (this volume)).

In [10], 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 (right-ward) extraposition is illustrated. In short, comparatives such as comparative adjectives and the adverb *andere* as in the following example, license corresponding comparative phrases (such as phrases headed by *dan* (than)) by means of a feature which percolates according to the extraposition principle.

- (1) ...*niks anders* doen *dan* *almaar* ruw materiaal verzamelen
 ...nothing else do but always raw material collect
 (...do nothing else but collect raw material)



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 regering niets doet
 The question is justified why the 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 nooit dan gisteren
 Lower was the rate never than yesterday
 The rate was never 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 1. 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 1) which indicate for each node the begin and end position in the sentence respectively.

The following XPATH query enumerates all examples 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
```

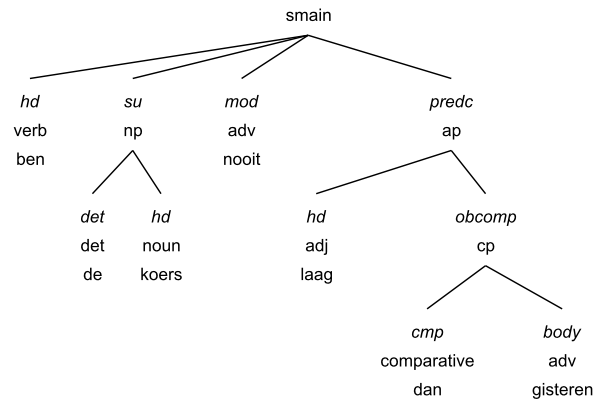


Figure 1: Dependency structure for *Lager was de koers nooit dan gisteren*

The query can be read as: find root sentences in which there is a daughter node, which 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 licensors 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 goed wordt als mijn forehand
 Further want I that my backhand just-as good becomes as my forehand
Furthermore, I want my backhand to become as good as my forehand

Some typical, genuine, examples are listed below. 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.

- (5) a. Liever betaalden werkgevers een (hoge) verzekeringspremie , dan opgescheept te zitten met niet volwaardig functionerende medewerkers (Algemeen Dagblad, January 15, 1999)
 b. Nooit eerder waren zoveel van de waterlelieschilderijen bijeen op één tentoonstelling als nu in Londen (Algemeen Dagblad, January 27, 1999)
 c. Liever huppelt ons land over het hoogpolige tapijt van de VN dan kennis te

nemen van de realiteit in het oerwoud en op de savanne van Afrika (Algemeen Dagblad, September 25, 1999)

- d. Beter is het te zorgen dat ziekenhuizen hun verplichtingen volgens de huidige BOPZ gaan nakomen, dan de rechten van patiënten nog verder aan te tasten (Algemeen Dagblad, August 18, 2001)
- e. Dus wat anders konden de LPF'ers de afgelopen week dan zich stil houden (Volkskrant June 1, 2002)
- f. Zozeer drukte de 300.000 mensen tellende Chinese gemeenschap haar stempel op de stad, dat zij de bijnaam 'Hongcouver' kreeg (NRC, August 23)
- g. Hetzelfde moet gebeuren als wanneer een man een mooie vrouw ziet, of andersom natuurlijk (Algemeen Dagblad, December 29, 2001)

The examples show that at least in some cases, the possibility of extraposition of comparative phrases from topic must be allowed.¹ More importantly for the purpose of this paper, we hope that the example illustrates that huge parsed corpora are helpful to find new empirical evidence for fairly complicated and subtle linguistic issues.

4 Learning Similar Words

The distributional similarity hypothesis is that words which occur in the same context have similar meanings [6]. In order to learn which words have similar meanings, concrete similarity measures have been applied to pairs of weighted context feature vectors, where each vector represents the context of a word, as observed in large text collections. An overview of such methods is given in [5]. The results have been shown to be practically useful for Question Answering, Information Extraction, Paraphrasing, etc.

Many of earlier distributional similarity methods used a *bag of words* to describe the context of words, but in more recent years, approaches where the context of a word is defined by reference to syntactic structure have become more common, and in many cases perform better than the earlier methods [9].

For Dutch, Van der Plas and Bouma [12, 11] compared the performance of various commonly used vector-based methods. In their approach, context features

¹The examples sometimes appear to indicate a somewhat old-fashioned, poetic style, perhaps explained by the following biblical citation, which also employs the syntactic phenomenon discussed here:

- (i) eerder gaat zoo'n kameel door het oog van een naald, dan dat een rijke in zou gaan in het koninkrijk der hemelen

for nouns are defined with respect to the syntactic dependency the noun occurs in. Typical features include 'occurs as the direct object of the verb *to paint*', 'occurs as an apposition to the noun *bag*' and 'occurs in a coordination with the noun *body*'. Each noun is represented by a vector, where each cell indicates the weighted score of a particular context feature. This score could be the raw frequency of the feature found in a large parsed corpus, but the authors show that pointwise mutual information [2] provides a more informative score.

Once all nouns are represented by vectors, a similarity measure is used to compute the similarity of two nouns. Van der Plas and Bouma use a particularly simple method, referred to as Dice-dagger, and introduced for a similar purpose in [3]. In order to compare two vectors v, w , Dice-dagger is defined as:

$$\sum_i 2 \cdot \frac{\min(v_i, w_i)}{v_i + w_i}$$

In a web-demo at <http://www.let.rug.nl/~gosse/SetsTwNC> the results of this method can be inspected. For this demo the TwNC newspaper corpus has been parsed with Alpino. For each word the demo returns those words which are closest with respect to the Dice-dagger metric. The method is best appreciated by looking at a number of examples. The method works very well for proper names:

Porsche Audi, Mercedes, BMW, Volkswagen, Volvo, Bentley, Renault, Skoda, Toyota, Saab, VW, Jaguar, Opel, Citroën, Fiat, Peugeot, Honda, Rolls-Royce, Bugatti

Tokyo Tokio, Montreal, Stockholm, Seoul, Kopenhagen, Boston, Melbourne, Genève, Buenos Aires, Zürich, Warschau, Oslo, Bangkok, Lissabon, Boedapest, San Francisco, Dresden, St. Petersburg, Praag

Groningen Eindhoven, Utrecht, Nijmegen, Zwolle, Tilburg, Den Bosch, Haarlem, Arnhem, Leiden, Breda, Leeuwarden, Almere, Amersfoort, Dordrecht, Delft, Maastricht, Lelystad, Enschede, Emmen

Tom Waits Patti Smith, Elvis Costello, Joni Mitchell, Bob Dylan, Brian Wilson, Lou Reed, Paul Simon, Randy Newman, Leonard Cohen, Nick Cave, Stevie Wonder, Frank Zappa, Mick Jagger, Captain Beefheart, David Bowie, Paul McCartney, Frank Boeijen, Van Morrison, Neil Young

Graham Greene Evelyn Waugh, Richard Klinkhamer, Tahar Ben Jelloun, Bram Stoker, David Yallop, Florian Illies, Italo Calvino, George Bernard Shaw, Bert Hiddema, E.M. Forster, Lu Xun, Elias Canetti, Stefan Heym, Gustave Flaubert, Frédéric Beigbeder, Michael Lewis, Anthony Powell, Albert Helman, Kingsley Amis

A number of examples with common nouns:

huis (house) woning, gebouw, pand, kamer, kantoor, boerderij, auto, villa, straat, winkel, tuin, flat, huisje, appartement, hotel, brug, paleis, kerk, ruimte (*house, building, house, room, office, farm, car, villa, street, shop, garden, flat, small house, apartment, hotel, bridge, palace, church, space*)

verliefdheid (enamour, love) heimwee, hartstocht, jaloezie, ontroering, erotiek, liefdesrelatie, fascinatie, vriendschap, begeerte, genegenheid, ontrouw, eenzaamheid, overspel, seksualiteit, gekte, verlegenheid, hysterie, waanzin, schuldgevoel (*homesickness, passion, jealousy, emotion, erotics, love relation, fascination, friendship, lust, love, infidelity, loneliness, infidelity, sexuality, madness, shyness, hysteria, madness, feelings of guilt*)

prei (leek) paprika, ui, bleekselderij, champignon, biet, spinazie, broccoli, selderij, kool, courgette, peterselie, bloemkool, knoflook, komkommer, venkel, tomaat, andijvie, aubergine, wortel (*pepper, onion, blanched celery, mushroom, beet, spinach, broccoli, celery, cabbage, zucchini, parsley, cauliflower, garlic, cucumber, fennel, tomato, endive, aubergine, carrot*)

In her thesis, Van der Plas evaluates the method using EuroWordNet as the target. If words are similar using the automatic method, then these words should also be similar in EuroWordNet. To compute similarity of words in EuroWordNet, the Wu and Palmer metric is used. In a first comparison, she shows that the syntax-based model performs much better than the word-based proximity model for this task.

She shows furthermore that it is important that huge annotated corpora are available, by comparing the result of an experiment which had access to 80 million words, with an experiment which had access to 500 million words. She provides scores for three classes of nouns, depending on frequency. For high frequency nouns the EuroWordNet score increases from .747 to .765; for middle frequency nouns the EuroWordNet score increases from .644 to .737. For low frequency nouns the increase in performance is largest: .488 vs. .666.

These experiments indicate not only that syntactically annotated corpora are an important resource for learning word similarity, but also that the size of such syntactically annotated corpora must be really huge if we also want to apply such techniques successfully for less frequently occurring words.

5 Learning Selection Restrictions

As a final example of the use of huge parsed corpora, we report on a study which shows that the incorporation of automatically learned selection restrictions im-

proves disambiguation performance of the Alpino parser.

Although parsing has improved enormously over the last few years, even the most successful parsers make very silly, sometimes embarrassing, mistakes. As a simple example consider the ambiguous Dutch sentence

- (6) Melk drinkt de baby niet
 Milk drinks the baby not
The baby doesn't drink milk

The disambiguation model of Alpino employs a wide variety of features. In particular, the model also includes features which encode whether or not the subject or the object is fronted in a parse. Since subjects, in general, are fronted much more frequently than objects, the model has learned to prefer readings in which the fronted constituent is analysed as the subject. Although the model also contains features to distinguish whether e.g. *milk* occurs as the subject or the object of *drink*, the model has not learned a preference for either of these features, since there were no sentences in the training data that involved both these two words.

In fact, in about 200 sentences of a parsed corpus of 27 million sentences *milk* is the head of the direct object of the verb *drink*. Suppose that an automatic learning procedure would need at least perhaps 5 to 10 sentences in order to be able to learn the specific preference between *milk* and *drink*. The implication is that we would need a (manually labeled!) training corpus of approximately 1 million sentences (20 million words). In contrast, the disambiguation model of the Dutch parser we are reporting on in this paper is trained on a manually labeled corpus of slightly over 7,000 sentences (145,000 words). It appears that semi-supervised or un-supervised methods are required here.

Note that the problem not only occurs for artificial examples such as (6); here are a few mis-parsed examples actually encountered:

- (7) a. Campari moet u gedronken hebben
 Campari must you drunk have
Campari must have drunk you / You must have drunk Campari
 b. De paus heeft tweehonderd daklozen te eten gehad
 The pope has two-hundred homeless-people to eat had
The pope had two hundred homeless people for dinner

In a recent paper, van Noord [14] describes a technique to include automatically learned lexical preferences in a maximum entropy model, using a method proposed in [8]. Preferences are expressed using pointwise mutual information association scores [4, 2]. The association scores are estimated using an automatically parsed corpus of 27 million sentences.

As illustration, consider the highest scoring verb/direct object pairs: *bijltje gooi_neer*, *duimschroef draai_aan*, *goes by time*, *kostje scharrel*, *peentje zweet*, *traantje pink_weg*, *boontje dop*, *centje verdien_bij*, *champagne_fles ontkurk*, *dorst les*, *fikkie stook*, *gal spuw*, *garen spin*, *geld_kraan draai_dicht*, *graantje pik_mee*, *krediet_kraan*, *draai_dicht*, *kruis_band scheur_af*, *kruit verschiet*, *olie_kraan draai_open*, *onderspit delf*, *oven_schaal vet_in*, *pijp_steel regen*,

We furthermore list the highest scoring objects of *drink*: *biertje*, *borreltje*, *glaasje*, *pilsje*, *pintje*, *pint*, *wijntje*, *alcohol*, *bier*, *borrel*, *cappuccino*, *champage*, *chocolademelk*, *cola*, *espresso*, *koffie*, *kopje*, *limonade*, *liter*, *pils*, *slok*, *vruchtensap*, *whisky*, *wodka*, *cocktail*, *drankje*, *druppel*, *frisdrank*, *glas*, *jenever*, *liter*, *melk*,

The paper argues that lexical affinities are also important for other types of dependency. Highest scoring pairs involving a verb and an adverbial are: *overlangs snijd_door*, *ten hele dwaal*, *welig tier*, *dunnetjes doe_over*, *omver kegel*, *on_zedelijk betast*, *stief_moederlijk bedeel*, *stierlijk verveel*, *straal loop_voorbij*, *uitein rafel*, *aaneen smeed*, *bestraf spreek_toe*, *cum laude studeer_af*, *deerlijk vergis*, *des te meer klem*, *door en door verrot*, *glad strijk_af*, *glazig fruit*,

The study provides experimental evidence that the inclusion of this type of lexical knowledge in the parser improves parsing accuracy. On the WR-P-P-H part of the D-Coi corpus (a set of manually verified dependency structures for 2267 sentences from the newspaper Trouw of 2001), the Alpino parser obtained a concept accuracy of 90.32%. After re-training the disambiguation model with the inclusion of lexical preferences, the accuracy went up to 90.73%. From this result, we are confident to conclude that huge parsed corpora are a useful resource. We can even use that resource to improve upon the parser that produced this resource in the first place!

Acknowledgement. The Lassy project is carried out within the STEVIN programme which is funded by the Dutch and Flemish governments <http://taal-unieversum.org/taal/technologie/stevin>

References

- [1] Gosse Bouma and Jennifer Spender. Gosse bouma and jennifer spender. the distribution of weak and strong object reflexives in dutch. In Frank van Eynde, Anette Frank, and Koenraad De Smedt, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*, Utrecht, 2009. LOT Netherlands Graduate School of Linguistics. This volume.

- [2] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.
- [3] James R. Curran and Marc Moens. Improvements in automatic thesaurus extraction. In *In Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59–67, 2002.
- [4] Robert Mario Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, 1961.
- [5] Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 107–114, Ann Arbor, June 2005.
- [6] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.
- [7] Heleen Hoekstra, Michael Moortgat, Bram Renmans, Machteld Schoupe, Ineke Schuurman, and Ton van der Wouden. *CGN Syntactische Annotatie*, December 2003.
- [8] Mark Johnson and Stefan Riezler. Exploiting auxiliary distributions in stochastic unification-based grammars. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 154–161, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [9] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.
- [10] Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374, 2002.
- [11] Lonneke van der Plas. *Automatic Lexicon-semantic Acquisition for Question Answering*. PhD thesis, University of Groningen, 2008.
- [12] Lonneke van der Plas and Gosse Bouma. Syntactic contexts for finding semantically similar words. In Ton van der Wouden, Michaela Poss, Hilke Reckman, and Crit Cremers, editors, *Computational Linguistics in the Netherlands 2004*, pages 173–186. LOT, 2005.
- [13] Gertjan van Noord. **At Last Parsing Is Now Operational**. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42, Leuven, 2006.

- [14] Gertjan van Noord. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the International Workshop on Parsing Technology (IWPT)*, ACL 2007 Workshop, pages 1–10, Prague, 2007. Association for Computational Linguistics, ACL.
- [15] Gertjan van Noord, Ineke Schuurman, and Vincent Vandeghinste. Syntactic annotation of large corpora in STEVIN. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.