Evaluating automatically annotated treebanks for linguistic research

Bloem, J.

Published in:
4th Workshop on Challenges in the Management of Large Corpora

4th Workshop on Challenges in the Management of Large Corpora

Workshop Programme

28 May 2016

14:00-16:00 – Session A

Introduction

Jochen Tiepmar,
CTS Text Miner – Text Mining Framework based on the Canonical Text Services Protocol

Jelke Bloem,
Evaluating Automatically Annotated Treebanks for Linguistic Research

Marcin Junczys-Dowmunt, Bruno Pouliquen and Christophe Mazenc,
COPPA V2.0: Corpus of Parallel Patent Applications. Building Large Parallel Corpora with GNU Make

Johannes Graën, Simon Clematide and Martin Volk,
Efficient Exploration of Translation Variants in Large Multiparallel Corpora using a Relational Database

16:00-16:30 Coffee break

16:30-18:00 – Session B

Adrien Barbaresi,
Collection and Indexation of Tweets with a Geographical Focus

Ruxandra Cosma, Dan Cristea, Marc Kupietz, Dan Tufiș and Andreas Witt,
DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva and Tsvetana Dimitrova,
Metadata Extraction, Representation and Management within the Bulgarian National Corpus

Closing Remarks

Evaluating Automatically Annotated Treebanks for Linguistic Research

Jelke Bloem
University of Amsterdam
1012 VB Amsterdam, Netherlands
j.bloem@uva.nl

Abstract

This study discusses evaluation methods for linguists to use when employing an automatically annotated treebank as a source of linguistic evidence. While treebanks are usually evaluated with a general measure over all the data, linguistic studies often focus on a particular construction or a group of structures. To judge the quality of linguistic evidence in this case, it would be beneficial to estimate annotation quality over all instances of a particular construction. I discuss the relative advantages and disadvantages of four approaches to this type of evaluation: manual evaluation of the results, manual evaluation of the text, falling back to simpler annotation and searching for particular instances of the construction. Furthermore, I illustrate the approaches using an example from Dutch linguistics, two-verb cluster constructions, and estimate precision and recall for this construction on a large automatically annotated treebank of Dutch. From this, I conclude that a combination of approaches on samples from the treebank can be used to estimate the accuracy of the annotation for the construction of interest. This allows researchers to make more definite linguistic claims on the basis of data from automatically annotated treebanks.

Keywords: Evaluation, Linguistics, Automatic Annotation

1. Introduction

This paper addresses an issue that is important for the application of large, automatically annotated corpora to linguistic research. A disadvantage of corpus-based methods in linguistics is that many phenomena of interest to theoretical linguists are used infrequently in naturalistic speech, and therefore are less likely to occur in smaller corpora. Automatically annotated linguistic resources contain the 'big data' that is necessary to study these phenomena empirically, but they inevitably contain errors as well, due to the imperfect natural language processing tools that were used to annotate them. As linguists are increasingly using large corpora as a source of empirical evidence, these issues have been acknowledged, but not explored systematically. In this work, I discuss four different approaches to evaluating data from automatically parsed corpora when a particular linguistic phenomenon is being studied. Current methods for evaluating the quality of annotation are too general for the purposes of studying a particular construction, and only measure the overall accuracy of the annotation of a corpus. While I use treebanks to illustrate the approaches because they seem to be the most common type of automatically annotated language resource, the approaches are also applicable to language resources created using other forms of automatic annotation, such as part-of-speech tagged historical texts or semantically parsed corpora.

1.1. Automatically Annotated Treebanks

Treebanks are text corpora that have been enriched with syntax trees or syntactic graphs (e.g. when dependency grammars are used), allowing linguists to search those texts for particular syntactic constructions and morphological features. Such queries will result in a list of only those constructions, which is much easier to study than an entire text. Due to advances in natural language processing, it has become possible to syntactically parse large amounts of text automatically, with fairly good accuracy. This has resulted in the creation of treebanks that are much larger than traditional, manually annotated corpora. From a linguistic perspective, the main advantage of these large-scale treebanks over manually annotated corpora is that they can be used to investigate rare constructions, co-occurrence patterns of uncommon words or small probabilistic effects. They also provide larger sample sizes or lists of examples of naturalistic data for more common linguistic phenomena.

Probabilistic effects in language have been discussed particularly in the study of alternations, i.e. multiple near-synonymous constructions that form two grammatical options for expressing the same meaning. Corpus studies of such phenomena have revealed that a number of factors from various domains of language (i.e. phonetics, semantics, pragmatics) may affect the choice between alternative constructions in such an alternation to varying degrees. The size of these effects can be interpreted as probabilities. The linguistic implications of the observation of such effects have been discussed by Bresnan (2007), who studied the English dative alternation. This multifactorial study experimentally tested whether probabilistic effects found in a previous corpus study corresponded to speaker preferences in a rating experiment between the two constructions of the dative alternation, i.e. 'I gave her the book' and 'I gave the book to her'. An earlier example of a multifactorial study of a linguistic optionality can be found in Gries (2001), who shows that many factors affect particle placement in English. In this study, Gries also computes effect sizes to quantify how influential the factors are, though the term 'probability' is not explicitly mentioned.

The main disadvantage of automatically annotated corpora is the error rate. While manually annotated or manually checked treebanks contain some errors, automatic annotation comes at the cost of annotation accuracy. The errors made by the parser will include systematic errors, where the parser has more difficulty with certain types of constructions than others. The parser may even be unable to annotate a particular construction correctly, thereby failing to provide the necessary means to search for it in a large corpus. Therefore,

when using automatically annotated treebanks for linguistic study, some sort of evaluation is necessary to make sure that the construction of interest was annotated correctly, or at least well enough for the purposes of the study. While the accuracy rate of the parser that was used to annotate the treebank is usually known, such a general measure of evaluation is not meaningful for most linguistic studies.

1.2. Construction-specific Querying

When studying a particular linguistic phenomenon or construction in a corpus, it may be more relevant to view the task as a form of information retrieval — all of the sentences instantiating the construction have to be retrieved from a larger data source (the corpus). The main difference with most information retrieval tasks is that the success of the search process depends on the quality of the annotation, rather than the quality of the search algorithm or the query. Nevertheless, I will use two basic measures from information retrieval, precision and recall, to illustrate the various approaches to evaluating the annotation quality of a corpus for a particular construction. In the context of this paper, **precision** is defined as the fraction of results from a corpus query that are instances of the construction that is being searched for, while **recall** is defined as the fraction of instances of the construction in the corpus that are retrieved. I will assume that corpus queries are perfectly written following the annotation format of the corpus being queried to specify exactly what the researcher wants to retrieve. Under this assumption, any imperfections in precision and recall occur only due to incorrect annotation. In reality, other issues may affect precision and recall as well, such as inaccurate formulation of the query or a lack of distinction between certain phenomena in the annotation format of the corpus. Since those would be problems of information retrieval rather than automatic annotation quality, I will not focus on them in this discussion.

2. Linguistic Studies Using Automatically Annotated Treebanks

Automatically annotated treebanks are a useful source of information in any study where large sample sizes are beneficial. Such treebanks have been made available for various languages. For Dutch, there is the 700 million word Lassy Large treebank (van Noord et al., 2013). For German, the 200 million word TüPP-D/Z (Müller, 2004) is available with automatic annotation. For English, the Google Books n-gram corpus (Lin et al., 2012) has been annotated syntactically, as well as the 4 billion word Gigaword v5 corpus (Napoles et al., 2012), to name but a few. Treebanks for a specific domain or language can be created as long as an automatic parser is available. This is the case for many major languages. Efforts have been made to make this technology more accessible to linguists who do not necessarily have a technical background, using techniques like example-based querying for treebanks (Augustinus et al., 2012), or systems where researchers can upload their own corpora to be automatically annotated, such as PaQu (Odijk, 2015).

These treebanks have already been used to study various linguistic phenomena. For Dutch, several applications of the Lassy Large treebank are discussed by van Noord and Bouma (2009). A study of extraposition of comparative objects by van der Beek et al. (2002) was used to illustrate the grammar used by the Alpino parser, but it attracted criticism for allowing too much extraposition. As evidenced by a note (van der Beek et al., 2002, 364, note 8), a reviewer claimed that such extraposition was not possible from the front of the sentence (the topic), however, a search of the large corpus revealed that such sentences were in fact being used in particular contexts. It was judged to be a probabilistic phenomenon, more or less acceptable depending on various factors. This shows that automatically annotated treebanks can also be used to refute claims based on linguistic theory. Bastiaanse and Bouma (2007) used syntactic structures from the treebank to argue that patients with Broca's aphasia have difficulty with constructions of higher linguistic complexity, rather than due to a frequency phenomenon. Bouma and Spenader (2008) studied the distribution of the Dutch reflexives *zich* and *zichzelf* with regards to the verbs with which they co-occur, where different verbs can select one or both of the options. These are examples of studying rare constructions or co-occurrence patterns, a task that automatically annotated treebanks are particularly suitable for.

Another such task is the study of probabilistic effects. Bloem et al. (2014) used a part of the Lassy Large treebank to study Dutch verb cluster constructions, a word order variation in which a variety of factors play a probabilistic role. Like English, the Dutch language may use auxiliary verbs to express features such as tense and aspect. In verb-final subordinate clauses, these verbs are grouped together at the end of the clause, and in main clauses, the first (finite) verb goes to the second position while the others are grouped together at the end of the clause. Interestingly, in subordinate clause two-verb clusters, both logical orders are grammatical:

(1) Ik zei dat ik het **gehoord heb**
I said that I it heard have
'I said that I have heard it.'

(2) Ik zei dat ik het **heb gehoord**
I said that I it have heard
'I said that I have heard it.'

Speakers may choose between the orders depending on a variety of factors relating to discourse, semantics, mode of communication or processing complexity (De Sutter, 2009; Bloem et al., 2014). Larger clusters of verbs are also possible, but not all of the logical orders are grammatical when three or more verbs are involved, although there is still variation.

Studying this phenomenon involves searching the treebank for groups of verbs in a particular syntactic configuration: an auxiliary or modal verb heading a participial or infinitival main verb. This study also replicates earlier work on a manually annotated corpus (De Sutter, 2009), showing that the errors caused by the automatic parsing are not necessarily a problem for linguistic study, although a few factors (i.e. word stress patterns) could not be studied due to the nature of the annotation that can be found in a treebank of written texts. While the manual study involved 2.390 instances of verb clusters, a sample of 411.623 clusters was gathered from the treebank.

Odijk (2015) showed that automatic annotation can even be used to study child utterances from the Dutch CHILDES corpus. This corpus was parsed with the Alpino parser for Dutch, even though this parser has not been trained on child language data. Spoken child language is a rather different domain than adult written language, making parsing errors likely. Nevertheless, a study of three near-synonymous Dutch degree modifiers that translate to 'very' was conducted on this data, along with an evaluation. The interesting thing about these modifiers is that two of them, *erg* and *zeer*, are used with adjectival, verbal and adpositional predicates, while one, *heel*, is only used with adjectival predicates. It is not clear how children acquire this difference. A corpus of child utterances and child-directed speech with syntactic information may reveal how much evidence there is for these constructions in child language. Useful results were obtained despite the issues, likely due to the focus on a particular linguistic phenomenon involving modifiers rather than larger syntactic structures — the two most common of the three degree modifiers were found with high accuracy.

Using the TüPP-D/Z treebank, auxiliary fronting in German three-verb clusters was studied (Hinrichs and Beck, 2013). Since three-verb clusters in subordinate clauses are a somewhat rare construction, and auxiliary fronting inside of them even more so, the massive size of the corpus was a requirement to be able to find enough instances. The authors observe what verbs participate in the construction and compare the treebank data to (much more sparse) information from diachronic corpora. For English, Lehmann and Schneider (2012) used a 580 million word dependency-parsed corpus to study the influence of specific lexical types on the English dative alternation. These types consist of 'triplets' of words: a ditransitive verb, a direct object head and an indirect object head — these slots are all filled with open-class words, requiring massive amounts of data to study.

3. Current Approaches to Evaluation

The quality of automatically annotated treebanks is usually evaluated by testing the performance of the parser that was used to create it. Therefore, treebanks are evaluated in the same way as parsers, using an overall accuracy score such as the word-based Attachment Score. This is the percentage of words that have been assigned the correct head in the syntactic structure (sentence-based variations or variations that include dependency labeling also exist). The Alpino parser (van Noord et al., 2006) that was used to create the Dutch Lassy Large corpus was evaluated using Concept Accuracy (the proportion of correct labeled dependencies) as a measure. A part of the corpus containing texts from various domains (e.g. books, newspaper texts) was manually verified in order to have a gold standard to compare against. This resulted in an accuracy score of 86.52%, but with clear variation across different domains (van Noord, 2009). Studies based on the corpus often report this score as a measure of quality.

However, even this is too general for the purposes of linguistic research. Rather than some domain of text, a researcher is primarily interested in one particular construction, and wants to know how accurately that particular construction was parsed in the corpus. If the parser often errs in labeling adjectives, this does not matter if one wants to investigate reflexives, but it would be a major problem for a study of adjectives. Parser errors cannot be entirely dismissed as random variation, some of the errors are likely to be systematic due to the nature of (most) syntactic parsing as a probabilistic task based on statistical learning.

One obvious consequence of this is that a parser is more likely to make mistakes when parsing rare phenomena, for which there was little evidence in the parser's training data. Phenomena that are of interest to linguists are often rare. Related to this is the fact that parsers generally perform worse on longer sentences, as shown in van Noord et al. (2006, 11, Fig. 5) for the Dutch Alpino parser. Sentence length is sometimes considered as a probabilistic processing effect in multifactorial linguistic studies, so this should also be considered. More errors occur when there is more ambiguity, regardless of whether the ambiguity is caused by semantic or syntactic factors. Multi-word units (idiomatic expressions) are also known to cause parsing errors (Nivre and Nilsson, 2004), but on the other hand, a parser may have been specifically improved to deal with multi-word units. Text types that are different than what the parser was trained on, such as the child utterances in the study by Odijk (2015), may also cause a higher error rate. Lastly, when the original text contains errors or unusual spelling, a parser is also likely to make more annotation errors.

Due to this possibility of systematic errors which may introduce more errors for certain constructions than for others, I believe that semi-automatic or manual construction-specific evaluation is necessary, using the knowledge of linguistic experts. Such an evaluation will provide insight into the quality of data gathered from automatically annotated treebanks for the purpose of linguistic study.

Some studies using automatically annotated treebanks have taken this approach. For example, Odijk (2015), in his CHILDES study, compares the parser accuracy for the specific words being studied against a manually annotated gold standard. However, such a gold standard is not always available, and the manual annotation was also found to contain errors. These errors were found by looking up the words manually, which is not possible if one is investigating more general constructions that can involve many words types, instead of particular words. Furthermore, the data set of child utterances of the constructions being investigated was fairly small, making a thorough manual evaluation more feasible than on large automatically annotated corpora. Therefore, this approach to evaluation is not always applicable. Bloem et al. (2014) took a semi-automatic approach by manually verifying a portion of the results of their syntactic queries for verb clusters. While the precision of the results can be measured with such an evaluation, it does not address the issue of recall. Any relevant construction that was annotated incorrectly and therefore missed by the querying procedure, will not be in the sample. In other studies, i.e. Hinrichs and Beck (2013), the paper does not address the issue of construction-specific evaluation at all.

In the next section I will discuss four possible approaches to construction-specific evaluation for linguistic research using an automatically annotated corpus. I will illustrate the four approaches with examples from the Dutch verb cluster

research described by Bloem et al. (2014), who conducted their research on the automatically annotated Lassy Large corpus. In listing these approaches. I am assuming that the linguist is faced with a corpus that is the end product of automatic annotation. This hypothetical researcher does not have access to, or is not able to use the tools that were used to annotate it, i.e. the methods do not require much technical knowledge. Without this restriction, other approaches could be taken, and have already been taken, such as re-training and/or evaluating the parser on an adapted text, using or creating a different annotation tool that is designed to target the construction of interest specifically, or simply parsing a large number of instances of the construction being studied and evaluating the parser's performance on that procedure. However, it is unlikely that someone whose main interest is linguistic research would have the knowledge or motivation to perform such procedures.

4. Linguistically Informed Evaluation

The main difficulty of this task, evaluating some subset of a large corpus (i.e. all verb clusters), is in gaining insight into precision and recall at the same time. The four approaches discussed here have various strengths and weaknesses relating to these two measures that I will discuss. An overview of the four approaches discussed in this section and their relative benefits is shown in table 1.

4.1. Manual Evaluation of the Results

The most obvious approach is a complete manual evaluation of the results by a linguist. This involves first formulating a query that matches a specific construction, and manually inspecting the results of the search. Any result that matched the query but was not actually an instance of the construction being studied, whether it was due to an annotation error or an imprecise query, is marked as false, and others as correct. A percentage can then be calculated, which represents the precision score of the query. However, this method has several disadvantages. Firstly, it may take a lot of time and resources to evaluate all results extracted from a large corpus in this way — Bloem et al. (2014) automatically extract 411.623 verb clusters from the 145 million word Wikipedia part of the Lassy Large corpus, too many to verify manually in any reasonable time frame. A representative sample of the results would have to be used. Secondly, this method may still miss constructions that were systematically misparsed. For example, if a researcher is searching for verb clusters but verbs in a particular type of cluster have been mistagged as adjectives, a search query for verb clusters will not find those mistagged instances, and the researcher will not know of their existence. The precision of the results can be measured with such an evaluation, but not the recall.

I have tested this method on the first 10.000 sentences of the Wikipedia section of the Lassy Large Large treebank using two-verb auxiliary clusters from subordinate clauses, as shown in (1) and (2) as an example construction. This sample of the corpus contains 193.378 tokens, covering 0.13% of the Wikipedia section of Lassy Large. A syntactic search for the target construction yielded 315 matching verb clusters, of which five were found to arguably constitute errors — these five verb clusters all had adjectival instead of verbal participles. An example of that would be 'He thought the door was closed', where 'closed' can be an adjective as well as a verb, and these five cases were annotated as adjectives. However, the fact that they could be verbs was also available in the annotation, so these five examples may not be errors depending on your theoretical perspective. Therefore, the precision of the annotation for this part of the corpus is $\frac{310}{315} = 0.984$, or 1. From this we can conclude that two-verb clusters were likely parsed with very high precision by the Alpino parser when this corpus was created.

4.2. Manual Evaluation of the Text

To solve the recall problem, it may be possible to do a manual evaluation of the text. By reading the original corpus text rather than just the results of a search query, even instances of the construction of interest that are completely misparsed can be found by the linguist. However, this is extremely labor-intensive — one will have to read a lot of text to find just one instance of a rare construction, even if only a part of the corpus is evaluated in this way. This takes away the main advantage of using a large automatically parsed treebank, and even if only a representative sample of the corpus is read, this is only feasible for common constructions. Therefore, I have chosen not to demonstrate this approach.

4.3. Fall Back to Simpler Annotation

Another solution is to fall back to a simpler annotation layer. It is generally the case that annotation of larger structures is more difficult. Lemmatizing and tagging (assigning a word class) only involves words, while parsing adds syntactic structure over multiple words. Queries based exclusively on word class will therefore result in fewer errors than searching on the basis of syntactic structure. For example, to retrieve verb cluster construction one would normally want to find a verb that is the head of another verb. But in this way, verbs that were attached incorrectly will erroneously be skipped. If the researcher simply searches for two verbs positioned next to each other in the linear order, these skipped verbs would also be included, at the cost of retrieving verbs that are next to each other coincidentally (i.e. as part of two different clauses) or as part of a larger structure. Comparing the result of the two procedures will produce a list of 'suspicious' instances, which can be evaluated manually (to be either included or excluded from the study) with less effort and better recall than when the results of a regular corpus query are manually evaluated. This does mean that the linguist will have to come up with some sort of word-class-based approximation of the construction under study using their knowledge of the language.

This approach is somewhat comparable to what the Sketch Engine does, as introduced by Kilgarriff et al. (2004). The Sketch Engine is a tool aimed specifically at lexicographers. It can extract collocation information and other information that is interesting for lexicography from a corpus, while ignoring other aspects of the annotation. It has been applied to a variety of corpora, including automatically annotated ones. However, it does make use of some syntactic structure annotation (which is not simple), namely to identify grammatical relations of collocations.

I have again tested this method on the first 10.000 sentences

Approach	Weaknesses	Strengths
Manual evaluation of the results	No recall, somewhat costly	Precision measure
Manual evaluation of the text	Extremely costly	Precision & recall
Fall back to simpler annotation	Misses POS-tagging errors	Recall measure
Search for particular instances	Hard to generalize result	Recall measure

Table 1: Overview of the strengths and weaknesses of each approach.

Error category	Frequency	Percentage
Part of longer cluster	56	74.67%
Parsing error	7	9.33%
Query error	12	16.00%
Total differences	75	100%

Table 2: Results of a comparison between syntactic search and POS-based search, listing the verb clusters found only by the latter one.

of the Wikipedia section of Lassy Large, comparing the result of a syntactic search with that of a part-of-speech (POS) based search using only the features of word class, lexical category and linear position in the sentence. The results of this are shown in table 2. I identified all results that were retrieved by the POS-based search but not by the syntactic search, and manually verified them. There were 75 such results in total. In 56 cases, the query had actually matched a group of two verbs that was part of a larger verb cluster of more than two verbs. Since only two-verb clusters, not three or four verb clusters are the target construction, these are not errors. It is difficult to avoid getting results from larger clusters when using POS-based search, since the difference is syntactic. In seven cases, there was an actual two-verb cluster that had been misparsed. These cases were mostly located in very long sentences with many parsing errors. The syntactic search had missed these, indicating a recall issue. A further 12 results also contained actual two-verb clusters and were annotated correctly, but were not identified by the syntactic search. This indicates a problem with the syntactic query rather than with the annotation. They can be considered retrieval errors, not annotation errors. Most of them involved verbal particles directly before or after the verb cluster, which I did not consider when formulating the syntactic query. Detecting such errors using this method can help the researcher to refine their queries. Overall, to the 315 verb clusters that were found in the previous section, we can now add $7 + 12$ new ones that were not identified by the syntactic search. This also allows me to compute the recall over this part of the corpus: $\frac{315}{334} = 0.943$, or $\frac{315}{322} = 0.978$ if we do not wish to consider the query errors as a recall problem. Again, this is only an estimation of recall, calculated over a fraction of the entire corpus and using manual comparison of the results. Furthermore, this estimate does not take into account that the part-of-speech tagging may also contain errors. Automatically annotated part-of-speech tags are not perfect either, they are just more correct than the parse trees. The final approach I will discuss does not make this assumption, but instead circumvents the annotation as much as possible.

4.4. Search for Particular Instances

It is possible to search for particular instances (types) of the construction without relying on the annotation at all. The linguist can choose some representative instances of the construction and search for it directly. For example, when searching for Dutch verb clusters, one could simply search the corpus for the string *hebben gehad* 'have had', one of many possible combinations of verbs. I will call this a 'string query', as opposed to a 'syntactic query' that one would normally perform on a treebank. This will only result in a limited number of results, but in a large automatically annotated corpus there can still be many results for a specific combination of words, even if the total number of instances of the general construction (verb-verb combinations in this case) is much larger. This string query does not rely on any syntactic annotation, and it would therefore find the verbs even if they were annotated completely erroneously, i.e. as a preposition heading a preposition. These results for the string *hebben gehad* can then be compared to results for the syntactic structure of 'hebben gehad' with these particular words to see whether there is a recall problem: if an example occurs in the string query but not in the syntactic query, it is annotated incorrectly in a way that makes it impossible to find with a syntactic query. If the researcher does this for various instances of the construction, they should get a clear idea of the reliability of the annotation and what sort of errors to look out for. However, it would be impossible to find all annotation errors in this way, since for most research questions it would not be possible to search for every instantiation of a construction.

One disadvantage of this method is that it requires generalization. If you measure the recall for the *hebben gehad* verbal cluster, you might assume that the recall is similar for other two-verb clusters, but this is not necessarily true — perhaps the recall is worse for less frequent words. This concern may be alleviated by sampling a variety of instantiations of the construction. An advantage of the method is that it can be used not only to evaluate the quality of the automatic annotation, but also of the annotation scheme. It has been argued, most notably by Sinclair (2004), that it is better to avoid any sort of annotation if possible, as this already imposes theory upon the data. It may be the case that the annotation scheme of the corpus makes incorrect theoretical assumptions, groups different phenomena together into one category or makes arbitrary distinctions. By performing a query that avoids the annotation altogether and

comparing its results to those of a query that does make use of the annotation, such issues can be detected by a linguistic expert.

I have also applied this method to the Wikipedia data. I could not use only the first 10.000 sentences of the corpus, because there is only one verb cluster instance that occurs more than once in this sample. Instead, I took the first 300.000 sentences of the corpus and searched for the string *hebben gehad*, a common combination of common words, as well as the syntactic version: the verb cluster *hebben gehad*. The syntactic search resulted in four correct examples of *hebben gehad* verb clusters, while the string search provided 14 results, of course including the 4 correct examples from the syntactic search. Nine of the other results were actually verb clusters in main clauses, which are not the target construction, but the distinction cannot be made with just a string search because main clause verb clusters and subordinate clause verb clusters have the same form. However, the remaining string search result was indeed a valid *hebben gehad* verb cluster. On closer examination it had been parsed incorrectly, and therefore it could not have been identified by the syntactic search. It occurs in a sentence with an unusual structure that the parser apparently failed to parse completely, with the main verb *hebben* being left outside of the sentence structure. From this string search, it appears that there were actually five clusters, of which four were identified by the syntactic search. The recall here is 0.8, over this extremely limited sample for this construction.

5. Conclusion

In this paper, I have discusses the issue of using and evaluating linguistic data from automatically annotated treebanks for the purposes of linguistic research. I compared four approaches to evaluation and illustrated them with examples based on a recent linguistic study. These evaluation methods may help to alleviate the concerns that linguists often have about the inaccuracies of such corpora and provide more detail than traditional measures of parsing accuracy when the goal is to study specific constructions.

Since the proposed methods all have different advantages and disadvantages, it would be best to combine them when studying a particular construction. Manual evaluation of the results can be used to determine the precision of a corpus query's results, while searching for particular instances can be used to calculate recall over a portion of the data, determining how many examples might have been missed. Falling back to simpler annotation can be used as a verification of the syntactic annotation of the corpus, even over larger amounts of data, and provide a rough estimate of recall.

While the methods do require some manual annotation effort, they allow a linguistic researcher to get a clearer impression of the quality of the annotation of the particular construction they are investigating in the corpus, while still preserving the advantage of being able to obtain many exemplars with relatively little manual effort. Reporting on such a construction-specific evaluation in a large-scale corpus or treebank study makes the results easier to interpret for those who are not familiar with the errors that an automatic syntactic parser might make. Clearer quantification of the error rate for the linguistic phenomenon that is being studied will also allow researchers to make more definite claims on the basis of data from automatically annotated treebanks. In future work, a larger-scale empirical evaluation of these approaches on a wider variety of constructions and corpora could be conducted to assess them in more detail, and perhaps to create a better reason for linguists to use automatically annotated treebanks in their studies. Furthermore, it may be interesting to investigate whether construction-specific accuracy scores can be incorporated into corpus-based statistical models of language phenomena as part of the margin of error.

Acknowledgements

I would like to thank an anonymous reviewer for their helpful suggestions, and Arjen Versloot and Fred Weerman for the insightful discussions on linguistics and corpora.

6. References

Augustinus, L., Vandeghinste, V., and Van Eynde, F. (2012). Example-based treebank querying. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, pp. 3161–3167.

Bastiaanse, R. and Bouma, G. (2007). Frequency and linguistic complexity in agrammatic speech production. *Brain and Language*, 103(1): pp. 78–79.

Bloem, J., Versloot, A., and Weerman, F. (2014). Applying automatically parsed corpora to the study of language variation. In Jan Hajic et al., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1974–1984, Dublin, August. Dublin City University and Association for Computational Linguistics.

Bouma, G. and Spenader, J. (2008). The distribution of weak and strong object reflexives in Dutch. *LOT Occasional Series*, 12: pp. 103–114.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in search of its evidential base*, pp. 75–96.

De Sutter, G. (2009). Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. In A Dufter, et al., editors, *Describing and Modeling Variation in Grammar*, pp. 225–255. Walter De Gruyter.

Gries, S. T. (2001). A multifactorial analysis of syntactic variation: Particle movement revisited. *Journal of quantitative linguistics*, 8(1): pp. 33–50.

Hinrichs, E. and Beck, K. (2013). Auxiliary fronting in German: A walk in the woods. In *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, p. 61.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105: pp. 116.

Lehmann, H. M. and Schneider, G. (2012). Syntactic variation and lexical preference in the dative-shift alternation. *Language and Computers*, 75(1): pp. 65–75.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the

Google Books Ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pp. 169–174. Association for Computational Linguistics.

Müller, F. H. (2004). Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). In *Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen*, volume 28.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 95–100. Association for Computational Linguistics.

Nivre, J. and Nilsson, J. (2004). Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Odijk, J. (2015). Linguistic research with PaQu. *Computational Linguistics in The Netherlands journal*, 5: pp. 3–14.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

van der Beek, L., Bouma, G., and van Noord, G. (2002). Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 2002: pp. 353–374.

van Noord, G. and Bouma, G. (2009). Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pp. 33–39. Association for Computational Linguistics.

van Noord, G., Mertens, P., Fairon, C., Dister, A., and Watrin, P. (2006). At Last Parsing Is Now Operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42. Leuven University Press.

van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., Linde, J., Schuurman, I., Sang, E. T. K., and Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns et al., editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pp. 147–164. Springer Berlin.

van Noord, G. (2009). Huge parsed corpora in LASSY. In F. Van Eynde, et al., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, volume 12, pp. 115–126. LOT.