Using automatically annotated corpora in language variation research

Jelke Bloem, Arjen Versloot, Fred Weerman
University of Amsterdam
1012 VB Amsterdam, Netherlands
{j.bloem, a.p.versloot, f.p.weerman}@uva.nl

Keywords: automatically annotated corpora, language variation, quantitative methods

1. Introduction

Due to advances in computational linguistics and natural language processing, it has become possible to annotate large amounts of linguistic data automatically, with fairly good accuracy. This has resulted in the creation of large corpora of texts enriched with, for example, syntactic information. These corpora are much larger than traditional, manually annotated ones and are then often used by computational linguists to test and further develop other natural language processing tools. However, there have not been many studies from other areas of linguistics making use of this type of resource. We provide a case study in the domain of language variation, a domain in which quantitative and corpus methods have been particularly successful. We replicated and extended a previous quantitative variation study that used manually and semi-automatically annotated data, by making use of a large automatically annotated corpus for Dutch. The larger scale of our investigation allows us to generalize the claims of the previous study to Dutch two-verb cluster variation in general.

We used the Lassy Large corpus as our source of automatically annotated Dutch language data (van Noord et al., 2013), currently the largest corpus for Dutch. It contains texts from various written sources annotated with full syntactic dependency trees. The sentences have been parsed automatically by the Alpino parser for Dutch (van Noord and others, 2006). This parser is currently the state of the art, and an evaluation over different types of text shows an average concept accuracy (in terms of correct named dependencies) of 86.52% (van Noord, 2009). Furthermore, a variety of tools are available to access this corpus, including GrETEL (Augustinus et al., 2012), a query engine developed in the CLARIN-NL Nederbooms project which allows linguists to search using natural language examples rather than having to learn a specific querying language. This querying-by-example method is more intuitive to work with than formulating syntactic queries, making the data more easily accessible for people without specialized technical skills. The trade-off is that not every aspect of syntax can be captured by example sentences and therefore not everything can be searched for, for example recursive structures. So far, this tool is only available for a limited part of the corpus, but it is currently in the process of being extended.

2. Verbal clusters

In our case study, we have examined order variation in Dutch two-verb clusters, which allow for a lot of optionality without a clear meaning difference. In such clusters, the auxiliary head can be positioned before or after the main verb:

(1) Ik denk dat ik het **begrepen heb**
I think that I it understood have
'I think that I have understood it.'

(2) Ik denk dat ik het **heb begrepen**
I think that I it have understood

In example 1 the head comes last, and we will call it the 2-1 order, example 2 is the 1-2 order. This phenomenon has been widely studied in Dutch syntactic literature (i.e. Evers (1975), De Sutter (2005), Coussé (2008), Arfs (2007)), and there already exists a manual quantitative corpus study to compare to (De Sutter, 2009). This research has shown that this variation is influenced by a wide range of variables ranging from syntactic to semantic to discourse phenomena and is best studied using a multivariate model as in (De Sutter, 2009).

3. Using automatically annotated data

Corpora that have been annotated by an automatic parser allow for more instances of particular linguistic constructions to be found. This does come at the cost of accuracy, since automatic parsers are not perfect. It can be expected that the larger sample size makes up for random parsing errors, however, systematic errors may skew the results. Care should be taken that the parser is able to actually annotate the constructions under investigation.

Larger sample sizes are especially useful for the study of language variation, where it is often found that the phenomena depend on multiple variables that need to be modeled. Starting with Gries (2001), such models have been used to discover the size of the effect that each variable has on a particular linguistic variation. Using manual corpora, the number of constructions they investigate are limited, and these studies use a selection of sentences from a corpus, a process that may introduce subjectivity. We believe that these quantitative studies can be further improved by using data from automatically annotated corpora.

Using such corpora requires an exact definition of the constructions under investigation, which must be formulated in terms of the available annotation. For example, when defining a verbal cluster in 2-1 order, one must at least say that it involves two adjacent verbs in a subordinate clause, where the head verb comes after the other one. Using GrETEL, a researcher can do this simply by typing a sentence such as example 1, and specifying that the relevant properties are the part-of-speech of the two verbs, as well as their ordering. The system then automatically infers that it should look for any two such verbs that are positioned the same way in the syntactic tree as the verbal cluster of example 1. For examples of two other case studies using this tool, we refer to Augustinus et al. (2012). These two examples are about nominalization and particles in verbal clusters in Dutch. Only Dutch examples are available, because while the principle of example-based querying could be used in any language, the implementation of this tool is specific to the Lassy annotation format, which is specific to Dutch.

The tool still has some limitations, which prevented us from using it for this case study. At the time of writing, the search did not yet cover the entire corpus, and querying with an example sentence is not as powerful as using a query language, particularly when one wants to exclude words or parts-of-speech from occurring in the data. However, GrETEL does allow editing of the underlying search query, so that more technologically savvy linguists can get the data they want while still basing their search on an example sentence. An example of a linguistic study with a workflow that used both GrETEL and Dact, another tool for the Lassy corpora, can be found in Augustinus and Van Eynde (2012), though no automatically annotated part of the corpus was used.

A few linguistic studies that make use of an automatically parsed corpus can be found. Lehmann and Schneider (2012) used a 580 million word dependency-parsed corpus of English to study the dative alternation, a well-studied case of variation. They tested the effect of specific verbs and their arguments, therefore requiring vast amounts of data to get enough examples of each combination of words. The only such study that we are aware of for Dutch is on the optionality of the *om*-complementizer. Bouma (2013) created a multivariate regression model using data from part of the LASSY Large corpus, similar to the present study. He finds that this variation is best explained using variables related to semantics and processing complexity. He also reports on the predictive power of the model (a concordance score of 0.809).

4. Method and data

In our study, we aim to follow the methodology of De Sutter (2009) as closely as possible, but using the Lassy Large corpus, instead of manually annotated newspaper texts. We also take advantage of the abundance of data by extending the study to more types of verbal clusters. De Sutter only studied verbal clusters in complement clauses with the complementizer *dat*, containing a participle main verb (no infinitival clusters), and only clusters with the non-modal auxiliaries *hebben* ”to have”, *zijn* ”to be” and *worden* ”to be”. We also obtained data on other types of clauses (including main clauses), infinitival clusters, and clusters with modal auxiliaries, simply by defining some additional corpus queries.

We extracted verbal clusters from two distinct parts of the LASSY Large corpus. From the Wikipedia part, which consists of the entirety of the Dutch version of the freely editable online encyclopedia Wikipedia on the 4th of August, 2011 (about 145 million words) we extracted 411.623 two-verb verbal clusters, and from the Europarl part, which consists of the mostly translated proceedings of the European Parliament (37 million words) we extracted 467.521 verbal clusters. Using two subcorpora lets us check whether our findings generalize to another domain.

We then defined a multivariate logistic regression model, using the same predictor variables as De Sutter (2009) wherever possible. In total there were 9 variables in the original study (with two additional ones, CLAUSE TYPE and FINITENESS to distinguish our additional constructions), but we limit our discussion to the ones that we had to operationalize differently due to limitations imposed by the annotation scheme of the corpus.

5. Results

We can summarize the effects of the explanatory variables used in our regression model by stating that the effect sizes and directions are largely comparable to those found by De Sutter (2009), except in cases where variables had to be operationalized differently. In doing so, we were limited to the information available in the annotation of the Lassy Large corpus. We will proceed to discuss only these problematic cases. For other variables that are not mentioned here we confirmed the findings of de Sutter. For a full comparison between the studies, we refer to Bloem et al. (2014).

One explanatory variable could not be included in our model at all — DISTANCE BETWEEN ACCENTS, which relates to syllable accents, phonological information that is not included in the annotation. Another variable, SYNTACTIC PERSISTENCE, which measures whether a previous verbal cluster had the same order and could have had a priming effect, was not included because the Wikipedia texts from the corpus are edited by multiple authors at different times, meaning that the writer of one verbal cluster might not have read the previous.

Two other variables could be included, but had to be operationalized differently. EXTRAPOSED CONSTITUENTs are constituents that come after the verbal cluster, as in the example below:

(3) ... dat het **is** **gezegd** door de president
... that it **was said** by the president

Here, a prepositional phrase with an adjunct role follows the cluster, and it is syntactically attached to the cluster. These constituents after clusters have been hypothesized to relate to the verbal cluster order. De Sutter made a distinction between adjuncts and complements, but we could not extract this distinction from the corpus. We were still able to include the presence of these constituents, and whether they are syntactically attached to the verb or higher up in the tree. However, we found an association between extraposed constituents attached to the verb and the 1-2 order, while de Sutter found the opposite effect direction. This result is quite interesting, but unfortunately incomparable due to the different operationalizations.

To operationalize the TYPE OF THE AUXILIARY VERB, De Sutter divided the auxiliary verbs up into five grammatical classes: *zijn* "to be" as a copulative verb, *zijn* as a passive auxiliary, *zijn* or *hebben* "to have" as temporal auxiliaries, *worden* "to be", and unclassifiable. He developed an algorithm to identify them. It is described in De Sutter (2005, p. 205-230), involving 5 syntactic, 5 morphological and 2 semantic criteria. We did not try to re-create it because we would prefer to work with readily available corpus resources as much as possible for methodological demonstration purposes. We instead categorized the auxiliary verbs at the lexical level, simply distinguishing between the three words. As a result however, we found far smaller effects of the different categories than de Sutter did, they did not appear to make the most relevant distinctions.

Despite these issues, the other effects found by de Sutter were present in our data as well, and all of our findings were consistent across the two different subcorpora, Wikipedia and Europarl, indicating that these explanatory variables explain verbal cluster variation across different domains of text. Furthermore, we now also have data on main clause verbal clusters (found to be more associated with the 2-1 order, odds ratio = 0.34) and infinitival clusters (strongly associated with the 2-1 order, odds ratio = 0.03). We also tested the predictive power of our model and found it to have a concordance index of 0.8635 (after 100 bootstrap repetitions). De Sutter reports c = 0.803, though it should be noted that these two c-indexes cannot be directly compared between different models, since the variables are somewhat different. These values do indicate that the models are good enough for prediction tasks. It is clearly higher than the intercept of our model, which is 0.6035, and represents the odds of predicting an 1-2 outcome in the case where all the variables have their default value.

6. Discussion

We have shown that a previous investigation into Dutch verbal cluster variation, could also be carried out using exclusively automatically annotated corpus data. We were able to largely confirm the findings of De Sutter (2009), despite having slightly different models due to the annotation differences discussed in the previous section. Because this type of data gathering is less costly, we were also able to extend the study and show that the previously found explanatory variables also apply to other types of two-verb clusters, in two different domains of text.

This type of data, as found for Dutch in the Lassy Large corpus, seems to be suitable for linguistic studies in which a lot of data is necessary or helpful, and complements the recent trend of using large multivariate models. For the Dutch language, such data is becoming relatively easy to access using the CLARIN-NL GrETEL tool, which is still being developed further. We hope our results encourage other linguistic researchers to make use of such resources and test their hypotheses against large amounts of data.

Our case study presents some additional opportunities as well. We have yet to investigate clusters with more than two verbs, to which the automatic approach is uniquely suited. Larger verbal clusters are less frequent, and thus the best place to find rare constructions is in the largest available corpus. Now that large samples of data are easily available, it is also possible to explore the association between particular main verbs and the 1-2/2-1 order, providing more detail on possible semantic factors.

7. References

Arfs, M. (2007). *Rood of groen? De interne woordvolgorde in tweeledige werkwoordelijke eindgroepen met een voltooid deelwoord en een hulpwerkwoord in bijzinnen*. Göteborg University.

Augustinus, L. and Van Eynde, F. (2012). A treebank-based investigation of IPP-triggering verbs in Dutch. In *Proceedings of TLT*, volume 11, pages 7–12.

Augustinus, L., Vandeghinste, V., and Van Eynde, F. (2012). Example-based treebank querying. In *LREC*, pages 3161–3167. Citeseer.

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

Bouma, G. (2013). Om-omission in Dutch verbal complements. *Manuscript in preparation*.

Coussé, E. (2008). *Motivaties voor volgordevariatie. Een diachrone studie van werkwoordvolgorde in het Nederlands*. Universiteit Gent.

De Sutter, G. (2005). *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. University of Leuven: PhD thesis.

De Sutter, G. (2009). Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. *Describing and modeling variation in grammar*, 204:225–254.

Evers, A. (1975). *The transformational cycle in Dutch and German*, volume 75. Indiana University Linguistics Club Bloomington.

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

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

van Noord, G. et al. (2006). At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

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 Spyns, P. and Odijk, J., editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

van Noord, G. (2009). Huge parsed corpora in lassy. *Proceedings of TLT7. LOT, Groningen, The Netherlands*.