Chapter 6

Conclusions and Future Work

In this chapter, we first summarize our main conclusions and then point to some directions for future work.

6.1 Conclusions

This thesis presented four studies in Dutch syntax. In chapter 2, we saw that the Dutch it-cleft construction in fact consists of *two* distinct constructions. The first is analyzed as a transitive construction with a final relative clause. The subject *het* 'it' and the final clause map to the same f-structure, while the focused phrase functions as the non-subject argument of the copula. The second construction is analyzed as an intransitive construction with an expletive pronoun in subject position and a final complementizer clause. A total of three constituents map to the f-structure of the subject function: the expletive, the final phrase and the focused phrase. We were thus able to formulate accounts of both constructions which conform to the rules of canonical word order without violating the principle of subject-verb agreement.

In chapter 3 it was investigated if and how the factors that are claimed to influence the English dative alternation also influence the Dutch construction. In English, the two possible realizations differ with respect to both the order of the arguments and the syntactic category of the recipient. As a result, the literature on the dative alternation includes analyses in terms of both linearization constraints and NP or PP recipients preferences. In Dutch, word order and recipient category may vary independently. We hypothesized that we would find a differentiation between on the one hand general linearization constraints influencing the order alternations in Dutch, but not the NP/PP alternation, and on the other hand construction specific constraints which influence the NP/PP alternation. This hypothesis is partially borne out.

The verb lexeme only influences the syntactic category of the recipient: it may have a preference for a PP or an NP recipient, but not for a particular order. Pronoun type and definiteness in general were shown to influence argument order. So far the results were as predicted. But the contrast between pronouns and full NPs was also shown to influence the syntactic category. And most surprising, perhaps: it was shown that the classic linearization constraint on syntactic weight (light constituents precede heavier constituents) did not influence the order of the arguments in the midfield; it only has an effect on extraposition. The influence of these linguistic factors on the distribution of the alternants is in most cases probabilistic in nature: a factor may increase the chance of finding a certain realization, but does not lead to a categorical distinction between grammatical and ungrammatical. This poses a challenge for categorical models of language.

Chapter 4 shows that the syntactically marked combination of a preposition and a bare count noun (determinerless PP or PP-D) may be the result of various different syntactic constructions. These constructions differ in productivity and modifiability. We indicated how each of these constructions could be accounted for in a grammar, given the information about which preposition and which noun may participate in a PP-D and to what extend the combination allows modification. However, this information is generally not available. It is then shown that with the help of an automatically parsed corpus and various simple statistic measures, we can extract lists of PP-Ds of particular types and their modification potential semi-automatically. The quality of this extraction and classification method heavily depends on the availability of accurate noun countability information.

Chapter 5 focuses on the automatic classification of nouns according to their countability class(es). Following earlier work on English countability (Baldwin and Bond, 2003a,b), we were able to predict a Dutch noun's base countability class from its distribution in a corpus. While the English work focused on in-language learning, we experimented also with cross-lingual classification, using English training data to classify Dutch nouns. The best results, an accuracy of 85.7%, were achieved with a combination of both mono- and cross-lingual classifiers. Translation-based and transliteration-based classification proved to be remarkably accurate, but had very restricted coverage.

The translation-based results indicate that the countability for a given semantic concept is stable across languages. Based on this observation, we made an attempt at automatic noun countability classification using the semantic ontology EuroWordNet. The nouns were classified based on the known countabilities of the target word's synonyms, hypernyms, hyperonyms and cohypernyms. Again, we experimented with both mono-lingual and

cross-lingual classification. The results for the ontology-based classification methods are not as good as for corpus-based classification, with a maximum accuracy of 80.3%.

Although the topics and the methodology in each of the chapters varied widely, corpus data was involved in all chapters. It served as a source of examples and counterexamples in our investigation in the it-cleft constructions of Dutch and as the source of quantitative data in our probabilistic approach to the dative construction. In chapter 4, we extracted a repository of syntactically marked PPs semi-automatically from corpus data. Finally, we developed a set of corpus-based countability classifiers, which outperformed the ontology-based classifiers significantly. We thus showed that for both theoretical linguists and computational linguists, corpora provide valuable linguistic information.

In many cases, we depended on 1) large quantities of data and 2) syntactic annotation. Although there are some useful treebanks with manually edited syntactic trees, their size is limited. We therefore decided to complement this data with corpus data that was automatically annotated with dependency trees by the Alpino parser. We were thus able to find examples of rare types of it-clefts in Dutch. Since the various components of it-clefts (*it, to be*, and a relative clause) are very frequent, they do not make good key words for querying raw text corpora. Searching for those frequent (function) words will give a very large set of candidate clefts, almost all of which are false positives. But syntactic annotation helps to reduce the candidate set drastically. If one can specify the syntactic relation between those frequent words, one filters out many false hits, such as simple restrictive relative clauses, while keeping the clefts in the candidate set.

The syntactic annotation also helped us to identify sentences with a double object or dative PP construction. From these sentences we obtained the quantitative data necessary to identify the linguistic factors that influence the dative alternation. Again, it would have been impossible to identify double object constructions on the basis of POS-tag sequences or word strings, and manually annotated treebanks proved too small for certain queries.

Similarly, it was the syntactic annotation which allowed us to extract information about the verb heading a determinerless PP and its modification potential in chapter 4. Although with simple POS-tags it is possible to extract a large number of the PP-Ds, it would have been hard to extract information about the verbs and the modifiers (especially the postnominal modifiers). But information about the verb is crucial for separating the verbal PP complements from the independent PP-Ds. And information about the

modifiers a PP-D co-occurs with is necessary to determine the degree in which the PP is frozen.

Chapter 5 is the only chapter in which we did not use the full annotation that the Alpino parser produces. For countability classification, we only used chunk information. However, this chunk information was extracted from the automatically generated full parses. In short: we extracted valuable linguistic information from automatically parsed corpus data in each chapter of this thesis and we applied this information to four very different types of research into Dutch syntax. The parses that were automatically generated for the corpus sentences proved to be a close approximation of their actual syntactic structure, close enough to be useful for theoretical and computational linguistic research.

6.2 Future work

The research presented in this thesis provided answers to some linguistic questions but also raised other questions that had to be left for future investigations. Why is it that pronouns have such a strong preference for the subject position, leading to the two variants of transitive clefts? What semantic feature triggers a nominal to form a complementizer cleft instead of a relative clause cleft? Why is it that weight influences extraposition, but not the position in the midfield? Will a stochastic OT implementation predict the same frequencies for the various realizations of the dative alternation as we found in the corpus? How can we best model the chance of an optionally NP-D selecting preposition to occur in a determinerless PP? Is countability best modeled as categorical with some coercion possibilities, or is it better modeled as inherently gradient? Each of these are well worth further investigation.

But above all, this thesis shows the wealth of information that has become available with the development of accurate wide-coverage parsers. Automatically annotated data facilitates research that depends on syntactic annotation. Not only for topics that are extremely common—for these topics the small manually annotated corpora may suffice—but also for less frequent phenomena. Although the automatically annotated data will no doubt contain errors, and one has to be aware of the possibility of a systematic bias in the grammar, the large difference in size (compared to manually annotated corpora) and the relatively high quality of the annotation make it a very useful resource for future research.

This thesis only scratched the surface of the possibilities. Initiatives are

being taken to apply automatically parsed data for question answering,[1] but many more applications are possible. The parsed corpora can provide quantitative data about the conditions in which for example scrambling takes place, or about ordering differences between Flemish and Dutch, and they may facilitate the study of fairly infrequent constructions such as the Dutch dative passive. Well-established analyses and new hypotheses may be tested against large quantities of data, possibly revealing exceptions that have gone unnoticed.

Corpora form a natural source of data for linguistic research, and syntactic annotation enables the linguist to extract relevant information from this source. For this, linguists no longer have to rely solely on small scale, manually-annotated corpora: they can complement this data with large, automatically annotated corpora.

[1] `http://www.let.rug.nl/~gosse/Imix/`