Towards a Large Parallel Corpus of Cleft Constructions

Gerlof Bouma*, Lilja Øvrelid*, Jonas Kuhn†*

*Department of Linguistics
University of Potsdam, Germany
{bouma,ovrelid}@uni-potsdam.de

†Institute for Natural Language Processing (IMS)
University of Stuttgart, Germany
jonas.kuhn@ims.uni-stuttgart.de

Abstract

We present our efforts to create a large-scale, semi-automatically annotated parallel corpus of cleft constructions. The corpus is intended to reduce or make more effective the manual task of finding examples of clefts in a corpus. The corpus is being developed in the context of the Collaborative Research Centre SFB 632, which is a large, interdisciplinary research initiative to study information structure. We show how state-of-the-art NLP tools, like POS taggers and statistical dependency parsers, may facilitate powerful and precise searches, and we demonstrate through preliminary empirical findings how such a resource may provide new opportunities for the linguistic research of cleft constructions.

1. Introduction

Information structure studies the way in which the presentation of information is determined by the speaker's assumptions about the knowledge state of the hearer (Vallduví and Engdahl, 1996). Information structural concepts such as *topic* and *focus* have been claimed to show a clear correlation with certain syntactic constructions. Languages differ with respect to the extent to which they express information structure through syntactic structure. Scandinavian languages, for instance, have been argued to do so quite a lot, where certain syntactic constructions are employed to express information structural concepts such as focus (Gundel, 2006).

Cleft constructions have been widely studied within theoretical linguistics, partly for their role in structuring the information conveyed in an utterance in a range of different languages. In the example below, the choice to employ a cleft (1a) rather than a canonical clause (1b) may be influenced by the information status of the *clefted material* (here: *the young people*), as well as the *cleft clause* (*who are disappearing*). A syntactically similar alternation is found in a range of other languages, here exemplified by German in example (2), Dutch in example (3) and Swedish in example (4).

(1) a. It is [the young people] [who are disappearing].
b. The young people are disappearing.

(2) a. Es sind [die jungen Menschen], [die abwandern].
b. Die junge Menschen wandern ab.

(3) a. Het zijn [de jongeren] [die wegtrekken].
b. De jongeren trekken weg.

(4) a. Det är [ungdomarna] [som försvinner].
b. Ungdomarna försvinner.

The English cleft is claimed to focus attention on the clefted material (e.g., Prince (1978), Hedberg (2000)). The cleft clause typically contains presupposed or *known* (Prince, 1978) information, whereas the clefted material is new in some respects. Cross-linguistically, the information structuring properties of clefts may vary, however. In contrastive, empirical studies, this has been observed even for related languages from the Germanic and/or Romance groups (e.g., Dufter (2009), Gundel (2006), Johansson (2001)). These empirical studies of clefts have all included a considerable amount of manual effort and are hence naturally limited in scope: they use relatively small data sets and limit the number of languages involved. Gundel (2006) goes through an English-Norwegian translated novel by hand, whereas Johansson (2001) and Dufter (2009) employ parallel corpora to study the distribution of clefts contrastively, requiring a stage of time-consuming manual filtering following automatic searches.

Generally, linguistic phenomena within the realm of information structure are notoriously difficult to study using large-scale corpus-based methods. First, there are few resources which are annotated for information structure. Secondly, the creation of such resources by means of manual annotation is costly and has shown varied results in terms of annotator agreement (Ritz et al. (2008), for an overview). The generalization of such annotation by automatic means has furthermore shown little success. As a formally marked information structural device, the cleft construction provides a unique opportunity to study information structure on a large scale.

In this paper, we present our efforts to create a large-scale, semi-automatically annotated parallel corpus of clefts. The corpus is intended to reduce or make more effective the manual task of finding examples of clefts in a corpus. The corpus is being developed in the context of the Collaborative Research Centre SFB 632,[1] which is a large, interdisciplinary research initiative to study information structure. We will discuss how state-of-the-art NLP tools, like POS taggers and statistical dependency parsers, may facilitate

[1] http://www.sfb632.uni-potsdam.de

powerful and precise searches, and we demonstrate through a preliminary empirical investigation how such a resource may provide new opportunities for the linguistic research of cleft constructions.

2. The resource

In its current form the corpus is based on four languages from the Europarl corpus v3 (Koehn, 2005): Dutch, English, German and Swedish. Work is underway to add more languages, such as Greek and Spanish. The data has been retokenized, sentence aligned, POS tagged and parsed.

Retokenization A freely available toolchain (Procep) for retokenization of Europarl data has been developed during the creation of the cleft corpus.[2] The tools perform word- and sentence-level retokenization, taking into account language particular orthograhic conventions and abbreviations, and furthermore deal with some idiosyncracies of the Europarl data sets. The Procep toolchain also cleans up the raw data by converting remaining XML-entities to UTF-8, normalizing characters such as apostrophes, quotation marks, and hypens, etc., and in addition restructures some of the meta-data. The sentence boundary detection is performed using models trained through unsupervised machine learning with the NLTK Punkt Tokenizer package. The Dutch tokenization is handled by the tokenizer of the Alpino tools (van Noord, 2006).

POS tagging For German and English POS tagging, we used TreeTagger (Schmid, 1994). For Swedish, we employed MaltTagger (Hall, 2003). Both taggers were applied with standard pretrained models for the respective languages.[3]

Dependency parsing The English, German, and Swedish parts of the Europarl corpus were parsed with the freely available Maltparser (Nivre et al., 2006a), which is a language-independent system for data-driven dependency parsing.[4] MaltParser is based on a deterministic parsing strategy in combination with treebank-induced classifiers for predicting parse transitions. We trained the parsers on standard treebanks for these languages. The English training data set consists of the Wall Street Journal sections 2-24 of the Penn treebank (Marcus et al., 1993), converted to dependency format (Johansson and Nugues, 2007). The treebank data used for German and Swedish are the Tiger treebank (Brants et al., 2004) and the Talbanken05 treebank (Nivre et al., 2006b) respectively, and we employ the versions released with the CoNLL-X shared task on dependency parsing (Buchholz and Marsi, 2006). For English we use the parser settings from the English pretrained MaltParser-model available from http://maltparser.org. For German and Swedish, we use the learner and parser settings from the parser employed in the CoNLL-X shared task (Nivre et al., 2006c). The Dutch part of the corpus was analyzed with the wide-coverage Alpino parser (van Noord, 2006) and subsequently converted into dependency graphs.

For each of the languages, we have about 1.5M parsed sentences in dependency tree format. In terms of sentence alignment between pairs of languages, the average overlap between the languages lies above 80%.

3. Finding clefts

3.1. Syntax-based cleft extraction

English clefts, e.g. (1a), are relatively clearly marked by the lexical items that need to appear in a cleft and their rigid order. Extraction by regular expressions may therefore be a good choice (see Dufter (2009) for a study based on regular expression extraction followed by manual evaluation). Even in English, however, this approach has its limitations. For instance, English cleft clauses need not have a subjunction, as exemplified by (5a), and may also exhibit inversion of the expletive pronoun and copula, as in the yes-no question in (5b-c).[5]

(5) a. . . . and it is [this report] [I will be discussing on behalf of my group].
b. [Who] is it [who have to suffer]?
c. Is there no such will or is it [a sense of realism] [that is inducing us to refrain from tackling these issues and to leave the text as it is]?

Extracting such clefts with regular expressions will lead to very low precision. More problems arise for an approach based on regular expressions if we look at other languages: Word order variation may mean that expletive cleft-pronoun, copula, and clefted material occur in any order. For instance, in the Swedish examples in (6), the expletive pronoun follows the copula due to a clause-initial adverbial (6a), topicalized clefted material (6b) or an initial wh-clause in a yes-no question (6c). Like in English, Swedish clefts do not need to contain a subjunction. In the German example in (7), the expletive and the copula are not even adjacent, but separated by the cleft focus, because the verb is final in subordinate clauses. This situation also occurs in Dutch.

(6) a. Nu är det [ordförandeskapet och rådet]
Now is EXPL the chair and council
[som måste komma . . .].
that must come
b. [Vårt palestinska folk] är det [som
Our Palestinian people is EXPL that
drabbats hårdast i området].
were affected hardest in the area.
c. [Vilken lag] är det [som skall tillämpas]?
Which law is EXPL that shall be applied?

(7) Ich hoffe . . . dass es [gerade dieser Teil] ist,
I hope that EXPL precisely this part is
[der das tragende Element des
that the bearing element of the
Erweiterungsprozesses sein wird] . . . .
expansion process be will

[2] http://sourceforge.net/projects/procep/

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] http://maltparser.org

[5] All examples in the following sections are authentic and taken from the cleft corpus.

Dutch and German have the additional problem that the subjunctions are homographs of frequent items such as articles and demonstratives.

Morphological and syntactic information can help us overcome these issues. For instance, the cleft clause could be found by looking for subordinate clauses independent of a subjunction. And rather than specifying the copula and expletive by string positions, we can express their relation as a position-invariant dependency.

In each of the included languages, the syntax of clefts is similar enough to fit in a single abstract syntactic representation. This schema is given in Figure 1. For each language, the schema is supplemented with constraints on ordering and realization of its components. The schema captures two types of (analyses of) clefts (Hedberg, 2000): the cleft clause can be linked to the clefted material (*relative-clause-like*) or to the copula (*complement-clause-like*). The use of this schema in search queries furthermore enables coverage of the main types of analyses assigned to cleft structures by our parsers. We do not take position as to which of these analyses is correct. Rather, we note that, since the parsers we employ either do not know about clefts or only very rarely assign a cleft analysis, many clefts in the corpus end up having one of these two analyses. Recognizing both is thus a means of increasing coverage.

3.2. Process

We use Prolog to extract the clefts from the dependency parsed Europarl. As a general purpose programming language, Prolog does not restrict us in our query writing and corpus investigations. At the same time, because Prolog is very well suited for querying large databases in a (semi-) declarative manner, we can set up a fast and convenient linguistic corpus query environment with very little effort.

The specifications of the cleft-schema are implemented as Prolog predicates that define clefts in terms of dependency trees. In Figure 2 is an example query for the relative-clause-like clefts in Dutch. The predicates in the query are all intended to be read as VS(O) sentences, so that `has_lemma(N,L)` can be read as 'node N has lemma L'. We can recognize the dependency schema of Figure 1 if we follow the `is_under` relations. Extra constraints added to this query are a) that the copula is the finite verb,[6] b) the cleft clause follows all the other parts, c) there is an obligatory subjunction which heads the subordinate cleft clause, d) the copula does not have any verbal daughters (so as to exclude perfective auxiliaries), and e) the focus is not a conjunction (to prevent leaving the clause and entering a second conjunct). The query itself thus represents a mix of general facts about clefts in a language and particularities of the parser and its grammar.

In subfigures (b) and (c) of Figure 2, we see two matching sentences. The first is indeed a cleft. It is in a subordinate sentence which causes the expletive and the copula to be non-adjacent. The second is a structurally similar non-cleft: the pronoun *het* 'it' in this example is referential, and the indefinite NP with a relative clause attached is used as a predicate.

[6] This is not necessarily the case in Dutch, but constraining it to be so increases the quality of the results. In future work, we may be able to lift this requirement.

	Query performance		
Method	Precision	Recall	F-score
regex, narrow	21.9	47.8	30.1
regex, broad	11.1	88.6	19.7
syntax, gold standard	53.0	84.1	65.0
syntax, automatic	43.8	54.7	48.7

Table 1: Evaluation of the Swedish cleft queries on gold standard and automatically assigned dependency structures.

3.3. Cleft query evaluation

Explicit annotation of clefts is not common in treebanks. An evaluation of cleft identification is thus difficult. An exception is the Swedish *Talbanken05* treebank (Nivre et al. (2006b), and references therein), which contains 201 annotated clefts (almost 2% of all sentences). The head verb of a cleft clause is specially marked in the annotation. We have evaluated our Swedish queries against this resource to get an idea of the level of performance we may expect and to identify possible problems with our queries.

The results of the evaluation is in Table 1. We have given two regular expression-based methods as baselines. The first ('narrow') is a rather restricted query, more or less a direct translation of the query used for English in Dufter (2009).[7] It requires the copula to directly follow the expletive, and then the subjunction *som* and a verb following within a token window.[8] The second baseline ('broad') takes into consideration the remarks about Swedish at the beginning of this section, concerning word order and the optionality of a subjunction. It allows the adjacent expletive and copula in any order and a following verb within a wide token window.[9] As we can see, both regular expression baselines have a rather low precision. The broad baseline, unsurprisingly, combines this with a very high recall.

Against these baselines, we consider two evaluations of our Swedish cleft queries based on the schema in Figure 1. We start by using the gold standard structural annotation to (re-) identify clefts with our query. This query is clearly more effective than the baselines. Note in particular that the hit in recall compared to the broad regex baseline is only small. Since we will run our query on automatically parsed text, a relevant question is what the impact of parsing errors is on the query performance. The last row in Table 1 shows that

[7] Our regular expression-based queries differ in that we identify the verbal head of the cleft clause, since this is the element annotated in our gold standard. Due to this, we need to employ a POS tagger. Whereas Dufter allows for an unrestricted number of tokens between the copula verb and the relative pronoun, we restrict the token window to 20 tokens in the narrow query and 100 in the broad query.

[8] `"Det|det" "är|var" [],{0,20} "som" [],{0,20} [pos="verb"] within s`

[9] `("Det|det" "är|var" | "är|var" "det") [],{0,100} [pos="verb"] within s`

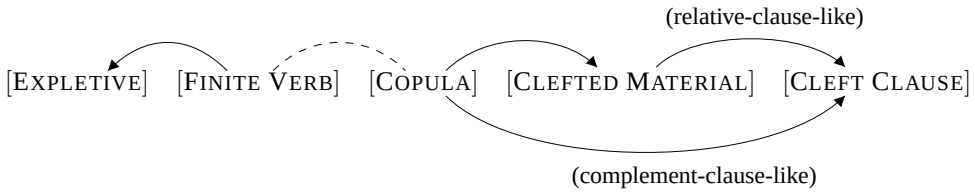

Figure 1: Abstract unordered dependency structure of cleft constructions.

(a)
```
contains_reltype_cleft(Sentence):-

        is_in_sentence(Expletive,Sentence)
        ∧ is_in_sentence(Copula,Sentence)              % four words in the
        ∧ is_in_sentence(CleftedMaterial,Sentence)     % same sentence
        ∧ is_in_sentence(Subjunction,Sentence)

        ∧ has_lemma(Expletive,'het')
        ∧ is_under(Expletive,Copula)                   % expletive directly under copula

        ∧ has_pos(Copula,'verb')
        ∧ has_lemma(Copula,'ben')                      % base form of 'to be'
        ∧ ¬ ( is_above(Copula,X) ∧ has_pos(X,'verb') )

        ∧ is_under(CleftedMaterial,Copula)
        ∧ ¬ has_posc(CleftedMaterial,'vg')             % 'vg' = conjunction

        ∧ is_under(Subjunction,CleftedMaterial)
        ∧ ( has_lemma(Subjunction,'dat') ∨ has_lemma(Subjunction,'die') )

        ∧ precedes(CleftedMaterial,Subjunction)
        ∧ precedes(Copula,Subjunction)
        ∧ precedes(Expletive,Subjunction)
        .
```

(b) … dat **het** nooit de gediscrimineerde **groepen** zelf **zijn die** hun standpunten naar voren kunnen brengen.
that EXPL never the discriminated groups self are that their opinions forward can bring
'…that it is never the groups that are being discriminated against themselves who can give their opinions.'

(c) **Het is** een **reeks** verslagen **die** wij ook in dit verslag hebben opgenomen.
it is a series questions that we also in this report have included
'It is a series of questions that we have also included in this report.'

Figure 2: Example Prolog query for Dutch clefts according to the relative-clause-like schema (a); dependency structure of a matching cleft (b); dependency structure of a matching non-cleft (c).

especially recall suffers from using automatic parses. We argue, however, that the recall/precision trade-off as compared to the regex baselines is still favourable. High precision means that a user of the cleft corpus has to discard fewer false positives. If a great quantity of clefts is desired, the loss in recall may be mitigated by the use of a large corpus, like Europarl. In this respect, it is important to emphasize that our syntactic queries themselves are designed to capture a broad variety of clefts. A lowered recall does therefore not necessarily indicate a systematic failure of the extraction method, as is the case for the narrow regular expression baseline.

Error analysis shows amongst other things that precision is affected by the presence of constructions built around a *referential* pronoun, a copula and a relative clause that are structurally invariant from clefts (8).

(8) Det är ett system som är känt över hela världen
it is a system that is known over whole world

Referential versus non-referential use of pronouns is not trivial to decide, even for humans. This is illustrated by the fact that we find the clause in (9) twice in the treebank, once occurring as a main clause and once as a subordinate clause, but with different annotations. The cleft annotation – glossed here – is arguably the correct one.

(9) Det är [bara två hälfter] [som kan utgöra en
EXPL is only two halves that can constitute a
enhet].
whole.

We also measured recall of the English query against the cleft dataset presented in Dufter (2009). The dataset was run through the parser and the query processor. Of the 459 cleft sentences, we recover 64.46%. Given the Swedish 53.73% recall when using automatically assigned syntactic structure, English cleft extraction seems to be a slightly easier task.

It is our intention that the cleft corpus may be used by third parties to investigate a wide variety of clefts in an efficient manner. The evaluation of our extraction setup for Swedish and comparison to regular expression based extraction suggests that our method of using an automatically parsed corpus and queries based on a cross-linguistic syntactic schema is a good way of achieving this goal.

4. A first look at the cleft corpus

In this final section, we shall cast a first look at the cleft corpus. This section fulfills the double function of giving a quantitative overview of the corpus and to demonstrate the kind of inquiries one could make using the corpus.

4.1. A cross-linguistic quantitative comparison

In Table 2 is a quantitative summary of the cleft corpus. In subtable (a), we give the sizes of each of the language subcorpora, in terms of words, sentences and extracted clefts. Subtables (b) and (c) summarize the overlap between cleft occurrences in language pairs.[10]

Even though the numbers first-and-foremost offer a quantitative overview of our extraction methods between languages, the comparison could also be linguistically meaningful, if only to suggest areas for further investigations in which the quality of the numbers is more carefully controlled. We begin by looking at the corpus sizes in subtable (a) in Table 2. For each language we have about 1.5M orthographic sentences, but the number of sentences that contain a cleft-like structure varies greatly between the languages. We find tens of thousands of possible clefts for English and Swedish, but German and Dutch clefts appear to be much rarer. Even correcting for the precision of the Swedish cleft queries (43%, Table 1), we can expect Swedish to have several times more clefts than either Dutch or German.

In subtable (b), we have calculated the probability of seeing a cleft in an aligned sentence given that we have a cleft in our source language. For English and Swedish, we can see that an estimated 29.0% of English clefts are translated into Swedish clefts (en→sv), but vice versa (sv→en) only 16.8%. This asymmetry is in the same direction as the one found by Johansson (2001) in his manual study. In spite of noisy data and low recall, the direction of the effect remains and indicates that the use of English cleft constructions is more constrained than in Swedish. Johansson (2001) notes that clefted elements in Swedish are more often anaphoric than in their English counterparts, and in particular, the clefted material is more often a personal or adverbial pronoun like *då* 'then', *därför* 'therefore' (10). With a corpus annotated for lemma information, POS tags and syntactic dependencies, one may operationalize such observations and test them against a considerably larger data set.

(10) Det är [därför] [jag förväntar mig att
EXPL is therefore I expect that
kommissionen ...]
the commission

Inspecting the sentences that are clefts in Swedish but not in English, we can find many types of alternative non-cleft constructions that are used instead. For instance, the Swedish-English pair below contains a cleft in Swedish, but a canonical structure in English (11).

(11) a. Vulnerable people find themselves hardest hit.
b. Det är [de sårbara människorna] [som
EXPL is the vulnerable people that
befinner sig i det hörn där motgångarna slår].
are in the corner where blows fall.

In his article, Johansson (2001) finds a correspondence between clefts in Swedish and so-called *reverse wh-clefts*[11] in

[10] Note that the figures are calculated from the automatic extraction and thus also reflect the errors in processing discussed in the previous section, for instance those due to parsing (affecting mostly recall), and the presence of referential copula-sentences (affecting precision). Thus, the term 'clefts' should be understood as 'cleft-like structures'.

[11] A reverse wh-cleft is a simple copula sentence with a postverbal free relative. These sentences do not fit in our current cleft-schemata because they lack an expletive.

(a)

	Corpus size		
Lang	Sents	Words	Clefts
de	1.5M	38M	2490
en	1.5M	40M	22060
nl	1.5M	37M	4545
sv	1.5M	33M	35680

Corpus size per language measured in sentences, words and cleft-like sentences.

(b)

	Target			
Source	de	en	nl	sv
de		30.7	19.1	43.7
en	3.4		6.1	29.0
nl	10.4	29.7		33.9
sv	2.8	16.8	3.9	

Conditional probability (%) of a cleft-like translation of a cleft-like sentence.

(c)

	Lang 2			
Lang 1	de	en	nl	sv
de				
en	19.6			
nl	60.3	19.0		
sv	16.7	10.8	12.8	

Ratio of observed overlap and overlap expected on the basis of translational independence.

Table 2: Corpus sizes and language comparison.

English. Casual inspection of the data indeed suggests that this pattern is common in our corpus, too. An example is in (12).

(12) a. This is what we are today asking the Commissioner for.
b. Det är [det] [vi i dag vill be herr
EXPL is that we today want ask Mr
kommissionären om].
Commissioner for.

The precise interpretation of such an asymmetry is a possible subject of future qualitative linguistic research. The cleft corpus is intended to greatly reduce the effort of contrastive study on clefts. By necessity, such an investigation will be of a different nature than Johansson's study, however. Johansson was able to do an exhaustive investigation of a much smaller corpus (although still a respectable 1M words for each of the two languages). With the Europarl-based cleft corpus, we can do a non-exhaustive (i.e., $<$100% recall) investigation of a much larger corpus ($>$30M words per language, and growing). Although it is much harder to do statistics in our situation, the researcher is in a position to get a wider qualitative view because the data set will contain more, and more varied, clefts in a larger number of languages.

The overlap between cleft occurrences in two languages can also be expressed by an association measure. In the table on the right hand side, we use the ratio of observed cleft co-occurrences and the expected cleft co-occurrences on the basis of independence between the two languages. Even though Dutch and German in general have few clefts when compared to English and Swedish, the overlap between the two languages is high: cleft co-occurrence is about 60 times higher than expected by chance. This might be related to the syntactic similarity of the two languages. In general, the table shows that the observed co-occurrences are many times higher than expected by chance. This suggests that cleft use, or at least, the use of cleft-like constructions, is correlated across the four Germanic languages that we have in the corpus thus far. It is our intention to investigate whether this correlation can be exploited to devise high precision extraction methods.

4.2. Cleft interpretation: Exhaustivity

After this rather high-level and quantitative look at the corpus, we briefly take up the more theoretical linguistic matter of exhaustivity in cleft interpretation, to illustrate how one might approach such an issue with the cleft corpus. English clefts are claimed to typically involve an expectation that the clefted material is the exhaustive list of elements for which the predicate realized in the cleft clause holds. For instance, the cleft *It's John and Bill that stole the cookies* also raises the expectation that nobody else did. Some support for this comes from the observation that inserting an additive particle leads to reduced acceptability (e.g., Krifka (2007); judgement from Drenhaus and Zimmermann (2009), who present a psycholinguistic investigation of exhaustiveness in German):

(13) *It's [John and Bill], too [that stole the cookies].

It is a topic of debate what the exact nature of this 'exhaustivity expectation' is, i.e., whether it is largely semantic or pragmatic, how strong it is, and to what extent it holds for other languages. This paper is not the place for an extensive literature review of these issues, but we do note that many arguments involve studying the interpretation of clefts containing negation, additive particles and focus particles. Finding these phenomena in a corpus is possible, as they can often be related to particular words or fixed expressions: *not* and negative quantifiers for negation; *too*, *also*, *amongst other things*, and perhaps even *first and foremost* for additive particles; *only* and *just* for focus particles, etc. Finding relevant, attested examples in their natural context is thus straightforward: we intersect the set of clefts in the corpus with the set of sentences containing lexical material of interest. The following example (14a) is a cleft from the Swedish part of the corpus, which contains an additive particle *även* 'also'. The aligned sentences in the other languages are also provided.

(14) a. Det är sålunda **även** [av förnuftiga och
EXPL is thus also of sensibility and
sakliga skäl] [som vi bör debattera
content reasons that we should debate
bägge programmen samtidigt].
both programmes together.

b. Es ist also **auch** [ein vernünftiger
EXPL is thus also a sensible
sachlicher Zusammenhang], [der uns rät,
content relation that us advises
die Debatte über beide Programme
the debates on both programmes
gemeinsam zu führen].
together IM lead.

c. The fact that the subjects are connected **also** suggests that we should hold the debate on both programmes together .

d. Er zijn goede inhoudelijke redenen om het
there are good content reasons IM the
debat over die twee programma's samen
debate about those two programmes together
te voeren .
IM lead.

Note in particular, that the German sentence (14b) is also a cleft containing an additive particle (*auch*). In the English sentence (14c), the additive particle is retained, but a canonical structure is used instead of a cleft. The Dutch version (14d) contains neither.

A more data-driven approach to the topic of cleft interpretation would be to treat clefts and particular lexical items as (anti-)collocations. For instance, if there is an incompatibility between a cleft and an additive particle, one would expect that additive particles occur considerably less often alongside clefted material than expected by chance. As an example of such an approach between languages, we could investigate whether English *only* occurs more frequently in sentences that are translational counterparts of cleft sentences. Such investigations often require vast amounts of data. Whether they are feasible with the cleft corpus remains a topic for future research.

5. Conclusion and future work

This paper has presented an effort to create a new linguistic resource, namely a large parallel corpus of cleft-like constructions. The phenomenon-specific, semi-automatic corpus annotation methodology we apply in this work adds a middleground between small hand-annotated resources and automatic general-purpose annotation of large resources. We believe this can be of high value for linguistic research. It may also provide datasets to prompt future specialized machine learning approaches.

Plans for future work include the application of our processing tools and query machinery to the newest version of the Europarl corpus (v5). As mentioned already, work is under way to add more languages to the corpus, specifically Spanish and Greek. We are furthermore in the process of making the corpus easily accessible through an interface based on the Open Corpus Workbench[12] and plan to make the resource publicly available within the near future.

Acknowledgments

The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in SFB 632 on Information Structure, project D4 (Methods for interactive linguistic corpus analysis).

The authors would like to thank Georg Jaehnig and Florian Marienfeld for their excellent work on the Procep tools for Europarl retokenization, as well as Andreas Dufter for sharing parts of his data set.

[12] `http://cwb.sourceforge.net/`

6. References

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2:597–620.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164.

Heiner Drenhaus and Malte Zimmermann. 2009. How exhaustive are you? Some ERP results on it-clefts, only-focus, and scalar implicatures. Talk presented at CSSP 2009, Paris.

Andreas Dufter. 2009. Clefting and discourse organization: Comparing germanic and romance. In *Focus and Background in Romance Languages*, Studies in Language Companion Seried 112. John Benjamins, Amsterdam.

Jeanette K. Gundel. 2006. Clefts in English and Norwegian: Some implications for the grammar-pragmatics interface. In V. Molnar and S. Winkler, editors, *Constrative Perspectives on Information Structure*. Mouton de Gruyter.

Johan Hall. 2003. A probabilistic part-of-speech tagger with suffix probabilities. Master's thesis, Växjö University, Sweden.

Nancy Hedberg. 2000. The referential status of clefts. *Language*, 76(4):891–920.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In Joakim Nivre, Heiki-Jaan Kaalep, and Mare Koit, editors, *Proceedings of NODALIDA 2007*, pages 105–112.

Mats Johansson. 2001. Clefts in contrast: a contrastive study of *clefts* and wh clefts in English and Swedish texts and translations. *Linguistics*, 39(3):547–582.

Philip Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit 2005*.

Manfred Krifka. 2007. Basic notions of information structure. In C. Féry, G. Fanselow, and M Krifka, editors, *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS) 6*, pages 13–56. Universitätsverlag Potsdam, Potsdam.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus for English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 2216–2219.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Joakim Nivre, Jens Nilsson, Johan Hall, Gülşen Eryiğit, and Svetoslav Marinov. 2006c. Labeled pseudo-projective dependency parsing with Support Vector Machines. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Ellen F. Prince. 1978. A comparison of *wh*-clefts and *it*-clefts in discourse. *Language*, 54:883–906.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Enric Vallduví and Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics*, 34:459–519.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.