Towards a large parallel corpus of clefts

Gerlof Bouma[†] Lilja Øvrelid[†] Jonas Kuhn[†‡]

[†] University of Potsdam, Dept. of Linguistics

[‡] University of Stuttgart, Insitute for Natural Language Processing (IMS)

May 21st, 2010

Motivation

- Information structural phenomena notoriously difficult to study using large-scale corpus-based methods.
  - few resources annotated for information structure
  - 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)
- As a formally marked information structural device, the cleft construction provides a unique opportunity to study information structure on a large scale.

Motivation

- Cleft constructions have been widely studied within theoretical linguistics
- Role in structuring the information conveyed in an utterance

(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.

Motivation

- Cleft constructions have been widely studied within theoretical linguistics
- Role in structuring the information conveyed in an utterance

(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.

- English cleft claimed to focus attention on clefted material (new), cleft clause is known

Motivation

- Cleft constructions have been widely studied within theoretical linguistics
- Role in structuring the information conveyed in an utterance

(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.

- English cleft claimed to focus attention on clefted material (new), cleft clause is known
- Property across languages?

Motivation

- We present our efforts to create a large-scale, semi-automatically annotated parallel corpus of clefts
  - Collaborative Research Centre SFB 632 – a large, interdisciplinary research initiative to study information structure

Motivation

- We present our efforts to create a large-scale, semi-automatically annotated parallel corpus of clefts
  - Collaborative Research Centre SFB 632 – a large, interdisciplinary research initiative to study information structure
- Intended to reduce or make more effective the manual task of finding examples of clefts in a corpus

Motivation

- We present our efforts to create a large-scale, semi-automatically annotated parallel corpus of clefts
  - Collaborative Research Centre SFB 632 – a large, interdisciplinary research initiative to study information structure
- Intended to reduce or make more effective the manual task of finding examples of clefts in a corpus
- Discuss how state-of-the-art NLP tools, like POS taggers and dependency parsers, may facilitate powerful and precise searches
- Enable contrastive, multilingual empirical investigations

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.

The Resource

- A freely available toolchain (Procep) for retokenization of Europarl data has been developed during the creation of the cleft corpus:
  - word- and sentence-level retokenization, taking into account language particular orthographic conventions and abbreviations

The Resource

- A freely available toolchain (Procep) for retokenization of Europarl data has been developed during the creation of the cleft corpus:
  - word- and sentence-level retokenization, taking into account language particular orthographic conventions and abbreviations
  - cleans up the raw data by converting remaining XML-entities to UTF-8, normalizing characters such as apostrophes, quotation marks, and hyphens, etc.

The Resource

- A freely available toolchain (Procep) for retokenization of Europarl data has been developed during the creation of the cleft corpus:
  - word- and sentence-level retokenization, taking into account language particular orthographic conventions and abbreviations
  - cleans up the raw data by converting remaining XML-entities to UTF-8, normalizing characters such as apostrophes, quotation marks, and hyphens, etc.
  - sentence boundary detection is performed using models trained through unsupervised machine learning with the NLTK Punkt Tokenizer package

The Resource

- A freely available toolchain (Procep) for retokenization of Europarl data has been developed during the creation of the cleft corpus:
  - word- and sentence-level retokenization, taking into account language particular orthographic conventions and abbreviations
  - cleans up the raw data by converting remaining XML-entities to UTF-8, normalizing characters such as apostrophes, quotation marks, and hyphens, etc.
  - sentence boundary detection is performed using models trained through unsupervised machine learning with the NLTK Punkt Tokenizer package
- For German and English POS tagging: TreeTagger (Schmid, 1994). For Swedish, we employed MaltTagger (Hall, 2003)

The Resource

- English, German, and Swedish parts of the Europarl corpus were parsed with the freely available MaltParser (Nivre et al., 2006), which is a language-independent system for data-driven dependency parsing.
- The Dutch part was analyzed with the wide-coverage Alpino parser (Noord, 2006) and converted into dependency graphs.

The Resource

- English, German, and Swedish parts of the Europarl corpus were parsed with the freely available MaltParser (Nivre et al., 2006), which is a language-independent system for data-driven dependency parsing.
- The Dutch part was analyzed with the wide-coverage Alpino parser (Noord, 2006) and converted into dependency graphs.
- For each of the languages, we have about 1.5M parsed sentences in dependency tree format
  Sentence alignment: the average overlap between the languages (pairs) lies above 80%

Syntax-based cleft extraction

- Limitations of regular expression-based approaches

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]?

Syntax-based cleft extraction

- Word order variation in other languages: expletive cleft-pronoun, copula, and clefted material in any order.

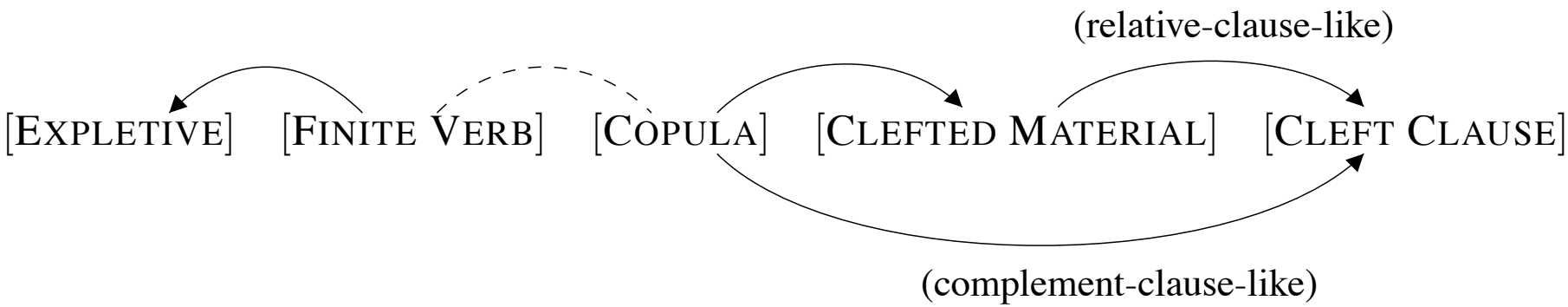
a. Nu är det [ordförandeskapet och rådet] [som måste komma . . .].
Now is expl the chair and council that must come

b. [Vilken lag] är det [som skall tillämpas]?
Which law is expl that shall be applied?

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

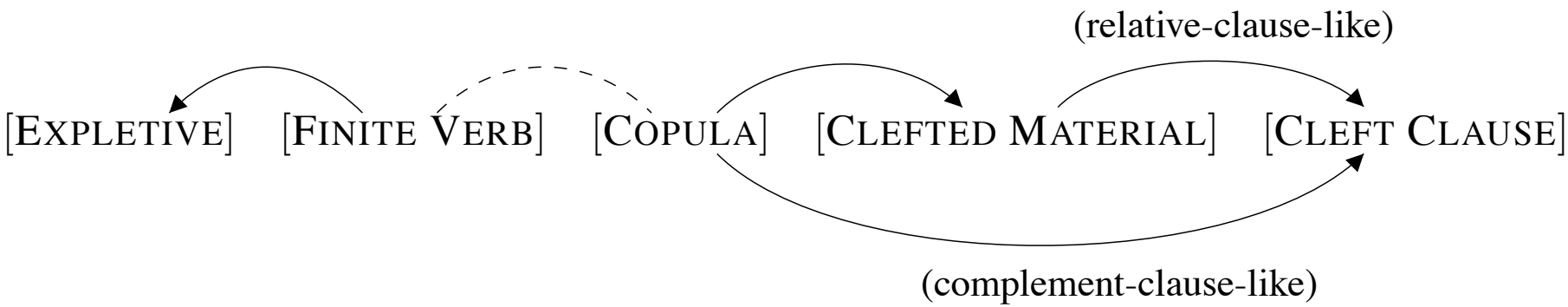
Syntax-based cleft extraction

- Morphological and syntactic information can help us overcome these issues
- Syntax of clefts is similar enough to fit in a single abstract syntactic representation



Syntax-based cleft extraction

- Morphological and syntactic information can help us overcome these issues
- Syntax of clefts is similar enough to fit in a single abstract syntactic representation



- Extraction with Prolog
- Predicates that define clefts in terms of dependency trees

Cleft query evaluation

- Cleft annotation in the Swedish *Talbanken05* treebank, contains 201 annotated clefts (almost 2% of all sentences)
- Two regular expression baselines, one strict (with *som*, fixed word order), one lax.

Cleft query evaluation

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

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

- Both regular expression baselines have a rather low precision. The broad baseline combines this with a very high recall.
- Gold query is clearly more effective than the baselines.

Cleft query evaluation

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

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

- Automatically parsed:
  - loss in recall may be mitigated by the use of a large corpus, like Europarl
  - our syntactic queries themselves are designed to capture a broad variety of clefts

Cleft query evaluation

- Referential pronouns

a. Det är [ett system] [som är känt över hela världen].
it is a system that is known over whole the world

Cleft query evaluation

- Referential pronouns

  a. Det är [ett system] [som är känt över hela världen].
  it is a system that is known over whole the world

- 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.

The corpus in numbers

	Corpus size		
Lang	Sents	Words	Clefts
de	1.5M	38M	2 490
en	1.5M	40M	22 060
nl	1.5M	37M	4 545
sv	1.5M	33M	35 680

Corpus size in sentences, words and extracted cleft-like structures.

The corpus in numbers

	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 seeing a cleft in an aligned target language sentence, given a cleft in the source language sentence.

The corpus in numbers

	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 seeing a cleft in an aligned target language sentence, given a cleft in the source language sentence.

German and Dutch cleft-likes are predicted much better from Dutch and German than from the other languages.

The corpus in numbers

	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 seeing a cleft in an aligned target language sentence, given a cleft in the source language sentence.

Swedish and English show an asymetry. Also observed in manual corpus study (Johansson, 2001).

Conclusion

- Presented a new linguistic resource
  - Large (multi-million)
  - Parallel (4 languages so far)
  - Syntactically annotated (PoS-tags, dependency parses)

Conclusion

- Presented a new linguistic resource
  - Large (multi-million)
  - Parallel (4 languages so far)
  - Syntactically annotated (PoS-tags, dependency parses)
- Future work: more languages, accessibility of corpus, a study into cleft exhaustivity using this corpus

References I

Dufter, A. (2009). Clefting and discourse organization: Comparing germanic and romance. In *Focus and background in romance languages.* Amsterdam: John Benjamins.

Hall, J. (2003). *A probabilistic part-of-speech tagger with suffix probabilities*. Unpublished master's thesis, Växjö University, Sweden.

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

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the mt summit 2005.*

Krifka, M. (2007). Basic notions of information structure. In C. Féry, G. Fanselow, & M. Krifka (Eds.), *Working papers of the sfb632, interdisciplinary studies on information structure (isis) 6* (pp. 13–56). Potsdam: Universitätsverlag Potsdam.

References II

Nivre, J., Hall, J., & Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (lrec)* (p. 2216-2219).

Noord, G. van. (2006). At Last Parsing Is Now Operational. In P. Mertens, C. Fairon, A. Dister, & P. Watrin (Eds.), *Taln06. verbum ex machina. actes de la 13e conference sur le traitement automatique des langues naturelles* (pp. 20–42).

Ritz, J., Dipper, S., & Götze, M. (2008). Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the International Conference on Language Resources and Evaluation (lrec).*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (p. 44-49).