Coreferenced corpora for information structure research

Erwin R. Komen[1]
Centre for Language Studies
Radboud University Nijmegen
E.Komen@Let.ru.nl

1 Introduction

1.1 Goal

Our project aims to find out how the *interplay* between *information ordering rules* and *syntax rules* has shaped the English language throughout its development.

- Research combines
  - Syntactic elements (phrases, hierarchy, NP type)
  - Information status (KNOWN, ASSUMED, NEW, etc.)
- Immediate goal
  - Information ordering rules for different stages of English
  - Interaction between syntax and information ordering rules
- Distant goal
  - Understand interaction between syntax and information ordering in general

1.2 Road map

Our project uses the following road map to gain understanding in the interplay between information ordering and syntax.

- Starting point
  - Syntactically annotated treebank texts
  - Available: Old English (YCOE), Middle English (PPCME2), early Modern English (PPCEME), late Modern English (PPCMBE)
- Add coreference information
  - Which constituent does each NP refer back/forward to?
  - Enriching manually (Cesac) vs. semi-automatically (Cesax)
  - Identify coreferential chains
- Interaction between syntax and information ordering (in general)
  - For each different time-period…
  - How is topic introduction marked?
  - How is an existing topic maintained?
  - How is topic-switch signalled?
  - How is new information focus expressed?
- Pilots (**this presentation**)
  - Average referential distance
  - Expression of "newness"
  - Subject referent switch

The reason to code only coreferentiality is that it provides a relatively objective measure, yet is powerful enough to provide the basic building blocks to derive information structural status (e.g. topic/focus, accessibility etc).

[1] Second affiliation: SIL-International, Erwin_Komen@SIL.org.

2 Add coreference information

2.1 Treebank to XML

- Approach
    - Convert corpora to XML
    - Stay close to existing standard TEI-P5
    - Add the information we want
    - Use existing Xquery as search engine within “CorpusStudio” (see **demo**!!)
- Our task is to enrich text with:
    - Pointer to antecedent
    - Feature: coreference type
        - NEW, IDENTITY, INFERRED, ASSUMED, …
    - Feature: NP type
        - PRONOUN, DEFINITE NP, INDEFINITE NP, QUANTIFIER, …
    - Feature: grammatical role
        - SUBJECT, ARGUMENT, PP-OBJECT, …
    - Extendable features!
- Advantages:
    - Features in general go with nodes in XML (either as attributes or as children of a particular tag-type)
    - Standard query language Xquery can be used and “fine-tuned” for work with corpora (e.g. using CorpusStudio).

2.2 Semi-automatic: Cesax

- Add information:
    - Develop own software: CESAX (Komen, 2011)
        - See website: http://erwinkomen.ruhosting.nl/software.
    - Semi-automatic coreference resolution
        - Manually adding (Cesac) → too time-intensive (Komen, 2009)
    - Improve algorithm as we work
- Coreference categories:
    - Referring: IDENTITY, CROSSSPEECH, INFERRED
    - Non-referring: NEW, NEWVAR, INERT
    - Additional: ASSUMED (=discourse-new, hearer-old)
- Cesax algorithm in a nutshell
    - Step 1: add features NPtype, GrRole, PGN
    - Step 2: resolve local coreference (partly already coded in treebank)
    - Step 3: mark particular NP types as “New”
    - Step 4: determine most likely candidate for antecedent of “next” NP based on a ranked set of constraints
    - Step 5: if the combination src-ant belongs to a “suspicious situation”, then consult the user, if not, automatically make the link
    - Step 6: allow the user to check the links that have been made automatically
    - Advantages:
        - Consistency: the same situations will be dealt with in the same way
        - Speed: sufficiently fast between requests to resolve ambiguities manually.
    - Performance:
        - Automatically done = 60% (6% of these are incorrect)
        - Remaining 40%: Suggestions are right in 50% of the situations

2.3 Where are we?

- Completely enriched texts (13):
    - Old English texts (Apollonius, St. Vincent)
    - Middle English texts (Sawles ward, Kent sermons, Horses)
    - Early Modern English (Fisher, Pinney, Behn)
    - Modern English (Brightland, Defoe, Skeavington, Long, Fleming)
- Partly enriched (6):
    - OE: Orosiu
    - ME: Capser, Polychronicon, Reynar, Wycliffe Sermons
    - eModE: Henry 4th history
- Work in progress…!

3 Coreferential chains

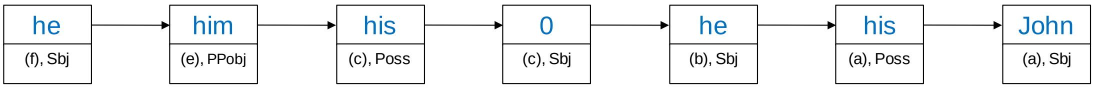
When a text has passed through Cesax, all coreference information is available, and can be combined into "coreferential chains".

3.1 What they are

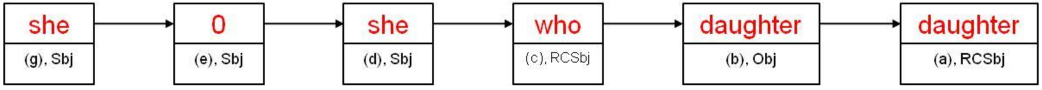
The concept of coreferential chains can be illustrated from the sample text in (1).

(1) a. [$_{sbj}$ John$_i$] entered the room where [his$_i$ daughter$_j$] watched television.
b. [$_{sbj}$ He$_i$] looked around,
c. and [$_{sbj}$ 0$_i$] saw his$_i$ daughter$_j$, [$_{sbj}$ who$_j$] sat on the couch.
d. [$_{sbj}$ She$_j$] looked up,
e. and [$_{sbj}$ 0$_j$] made a face at him$_i$,
f. while [$_{sbj}$ he$_i$] passed by.
g. [$_{sbj}$ She$_j$]'d had a rough day at school.

The coreferential chain of "John" runs backwards from line (f) until line (a):



The coreferential chain of "his daughter" runs from line (g) until line (a):



Real coreference chains can be found at http://erwinkomen.ruhosting.nl/results.

3.2 How we can make use of them

- Characteristics of coreferential chains
    - One chain = ordered list of constituents referring to one participant
    - Each element on the list contains features such as:
        - Person, number, gender of the participant
        - Text of the referring expression (e.g: "he")
        - Grammatical role (e.g: "object")
        - NP type (e.g: "pronoun")
        - The type of coreference relation (new, identity) it has with the antecedent
- Observation
    - Coreferential chains combine Syntax and Information State
- Imagine what we can learn from…
    - The distribution of chain lengths per genre or time-period
    - The percentage of particular NP types (Pro, Dem, …) on a chain

4 Syntax and information structure

This section looks at several test cases that are meant to illustrate the capabilities of the annotation scheme we use to answer questions that need syntactic information as well as information structure information. These test cases show that the enriched xml texts can effectively be queried using the existing Xquery standard to answer questions that need syntactic as well as information structure information.

Since corpus research that combines syntax with information state is a new research area, and since the number of texts available for this research is limited, the results presented are necessarily speculative.

4.1 Grammatical category and "newness"

The enriched corpora allow us to measure the relation between the grammatical category of an NP (e.g: subject, direct object, PP object) and the information state "new" (Komen, 2011).

- First step
  - What is "new"?
  - Algorithm to find "new"
- Definition of "new"
  - A constituent is referentially new if it
    - refers to a referent that has not been mentioned in prior discourse,
    - is not assumed to be known by the hearer,
    - does not contain an anchor to an established referent, and
    - can be referred back to in subsequent clauses.

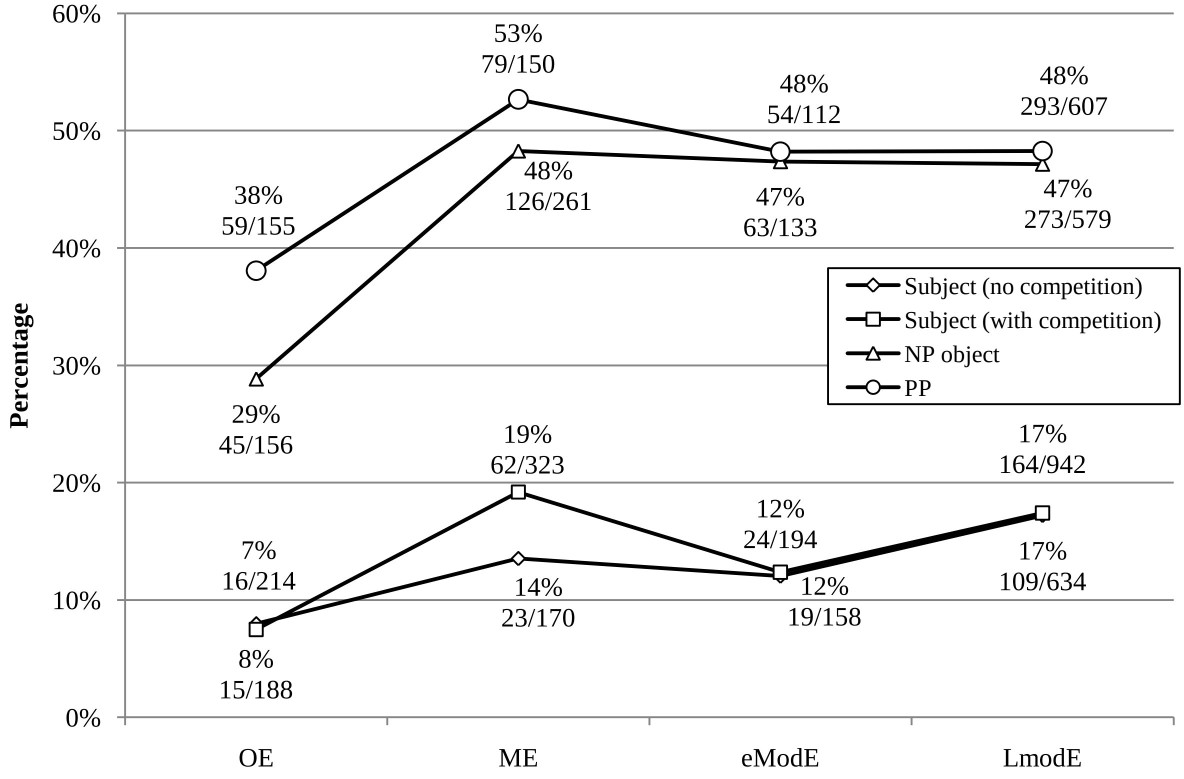


Figure 1 Percentage of "new" constituents per grammatical category

4.2 Subject referent switch

- Definition of Subject referent switch:
    - The number of times the referent of the subject switches,
      compared with the total amount of subjects in a text
- Restrictions:
    - Coordinate sentences have their own subject
    - Count the subjects of subordinate sentences or not?
    - Only look at 3rd person chains or not?
- Hypothesis:
    - *Since the preverbal field looses its unmarked linking function,
      subjects become the vehicle for unmarked linking as well as expressing the topic.
      This increased functionality should result in
      an increased switching between subject referents.*
- Results
    - Subject switching in main and subordinate clauses ("ipcls")
        - The general increase from OE to LmodE is expected
        - Unexpected is the stability between ME and LmodE
    - Switching of the referent of 3rd person main clause subjects
        - Increase from OE to LmodE is as expected
        - Relatively high number of switches in ME

	OE	**ME**	**eModE**	**LmodE**
Sbj, IP-Mat, chain=3rd person	527	527	363	1238
SubjectRefSwitch[ipmat3]	286	371	234	909
	54,3%	*70,4%*	*64,5%*	*73,4%*
Sbj, IP-Mat + IP-Sub	981	1158	840	2511
SubjectRefSwitch[ipcls]	591	846	603	1821
	60,2%	*73,1%*	*71,8%*	*72,5%*

Table 1 Subject Referent Switch

- Conclusions:
    - Need more data
    - Need to separate data according to genre

4.3 Average referential distance

- Referential distance (Givón, 1983):
    - "The amount of clauses between a Noun Phrase and its antecedent"
- Average referential distance (Givón)
    - The average of all referential distances
      for the NPs in a particular construction or word order.

- Test case: ARD of main clause subjects
    - Hypothesis:
        - *Since the preverbal field looses its unmarked linking function,*
          subjects become the vehicle for unmarked linking,
          which should result in a decrease in ARD for subjects in general.
    - Measuring the ARD for subjects
        - Select the subject nodes
        - Average their feature “IP-dist”
        - (Use “CorpusStudio” and Xquery on psdx files)
    - Results (see “AnySbj”):
        - Decrease from ME to LmodE is as expected
        - Small numbers in OE are unexpected

- Second test case: ARD of demonstratives
    - Hypothesis:
        - *The demonstrative system looses gender, the amount of demonstratives decreases, but there is no reason for them to change in referential distance.*
    - Distinguish between independent and dependent demonstratives
        - Dem (independent):
            - *this, that, those, these*
        - DemNP (dependent):
            - *this hat, that guy, those three trees, these men*
    - Results (see “Dem” and “DemNP”)
        - No big changes in “Dem” are as expected
        - Unexplained is the drastic decline for “DemNP”

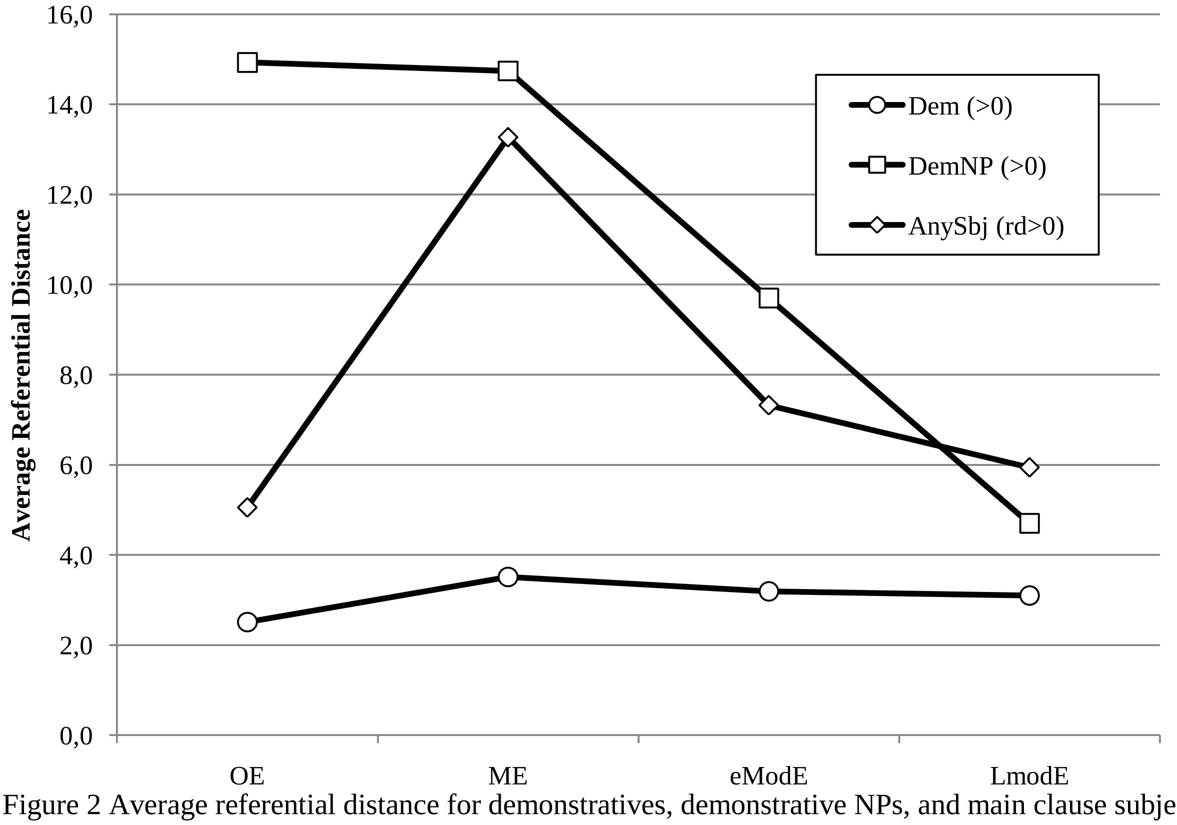


Figure 2 Average referential distance for demonstratives, demonstrative NPs, and main clause subjects

5 Discussion

- Corpus research combining syntax and information state
    - Need more enriched texts
    - Some results as expected:
        - Increased likelihood for NP objects to be referentially new
        - Subject switch increases over time
        - ARD of main clause subjects decreases from ME to LmodE
    - Some unexpected results
        - High subject switch in ME
        - Low main clause subject ARD in OE
        - Decreasing ARD of DemNP
- Annotation enrichment
    - Semi-automatic enriching using CESAX
    - Information State categories relatively objective (basic)
    - Extendable features for constituent nodes
    - Coreferenced corpora facilitate new research avenues

6 References

Givón, Talmy. 1983. “Topic continuity in discourse: an introduction”. *Topic continuity in discourse: a quantitative cross-language study*, ed. by Talmy Givón. Amsterdam, John Benjamins.

Komen, Erwin R. 2009. CESAC: Coreference Editor for Syntactically Annotated Corpora. In *7th York-Newcastle-Holland Symposium on the History of English Syntax (SHES7)*, 8. Nijmegen, CLS/Department ENglish Language and Culture: Radboud University.

Komen, Erwin R. 2011. “New changes in English - A diachronic perspective on the relation between newness and syntax”. *Linguistics in the Netherlands 2011 [AVT28]*, ed. by Rick Nouwen and Marion B. Elenbaas, 76-87. Amsterdam, John Benjamins.

Komen, Erwin R. 2011. Semi-automatic coreference enrichment for information structure research. In *Workshop on information structure and corpus annotation: theoretical and practical perspectives*. Oslo (Lysebu), Norway.