Bochumer Linguistische Arbeitsberichte 3

Beyond Semantics

Corpus-based Investigations of Pragmatic and Discourse Phenomena

Stefanie Dipper & Heike Zinsmeister (Eds.)

Bochumer Linguistische Arbeitsberichte

Herausgeber: Stefanie Dipper & Björn Rothstein

Die online publizierte Reihe "Bochumer Linguistische Arbeitsberichte" (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series "Bochumer Linguistische Arbeitsberichte" (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

© Das Copyright verbleibt bei den Autoren.

Band 3 (Februar 2011)

Gastherausgeber: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Heike Zinsmeister
Fachbereich Sprachwissenschaft
Universität Konstanz
Fach 185
78457 Konstanz

Erscheinungsjahr 2011
ISSN **2190-0949**

Stefanie Dipper & Heike Zinsmeister

Beyond Semantics:
Corpus-based Investigations of Pragmatic and Discourse Phenomena

Proceedings of the DGfS Workshop
Göttingen, February 23-25, 2011

2011

Bochumer Linguistische Arbeitsberichte
(Bla 3)

Table of Contents

Introduction

Stefanie Dipper[*] *and Heike Zinsmeister*[†]
[*]Ruhr-University Bochum and [†]University of Konstanz

This volume contains the papers presented at the Workshop *Beyond semantics: Corpus-based investigations of pragmatic and discourse phenomena*, which was organized as part of the Annual Conference of the German Linguistic Society (DGfS), held in Göttingen, Germany, February 23-25, 2011. The papers present corpus-based research on pragmatic and discourse-related phenomena. In recent years, focus of corpus-based research has moved from morpho-syntactic phenomena to semantics (e.g. word sense disambiguation, frame semantics, predicate-argument structure, temporal structure) and "beyond", i.e., to pragmatic and discourse-related phenomena (e.g. anaphora, information structure). In the latter field, it is often especially hard to transfer results from theoretical linguistics that are based on toy examples to naturally-occurring texts. Even provided explicit annotation guidelines, it is often difficult to annotate texts reliably.

The workshop brought together theoretical linguists who use texts and corpora for pragmatic or discourse-related research questions, and corpus linguists as well as computational linguists who create and annotate relevant corpus resources, or exploit them. The goal of the workshop was to enhance exchange between researchers of both fields, and thus to gain insight in the – possibly common – properties and peculiarities of the "beyond" phenomena.

We received 19 submissions from 8 countries, 13 have been accepted for presentation. The papers address the workshop topics from different points of view. A range of them deals with anaphoric relations and the way discourse referents are referred to. These papers relate corpus studies to linguistic models, such as the Givenness Hierarchy, Bridging Relations, Centering Theory, or Haspelmath's semantic maps (see the contributions by Chiarcos, Navarretta, Korzen & Buch-Kromann, Zulaica-Hernández & Gutiérrez-Rexach, and Aguilar-Guevara et al.). Presuppositional relations are sometimes also analyzed as a special type of anaphora; one paper presents research on automatic recognition of such relations (Tremper & Frank). Two papers address the annotation of information-structural features (topic and focus, see Riester & Baumann, Cook & Bildhauer). Several papers deal with the analysis of discourse-structure, either applying Rhetorical Structure Theory (van der Vliet et al.), or based on the Penn Discourse TreeBank (Versley), or from a conceptual and formal viewpoint (Buch-Kromann et al.). Two papers deal with the way speakers express their attitudes (Degaetano & Teich, Flöck).

In addition to the papers presented in this volume two invited speakers contributed to the workshop: Rebecca Passonneau (Columbia University) gave a talk on *Making Sense of Word Sense Variation*, Bonnie Webber (Edinburgh University) on *Patterns of Explicit and Implicit Clausal Connectors: What this might suggest for "beyond semantics"*.

We would like to thank the authors for their contributions as well as the members of the program committee. Special thanks go to our student assistants Christine Rieger and Melanie Seiss for help with formatting the volume. The workshop was in part supported by the German Research Foundation (DFG) and by Europäischer Sozialfonds Baden-Württemberg.

Workshop Organizers
Stefanie Dipper, Ruhr-University Bochum
Heike Zinsmeister, University of Konstanz

Keynote Speakers
Rebecca Passonneau, Columbia University
Bonnie Webber, University of Edinburgh

Program Committee
Maria Averintseva-Klisch, Tübingen University
Matthias Buch-Kromann, Copenhagen Business School
Philippa Cook, Freie Universität Berlin
Markus Egg, Humboldt-Universität zu Berlin
Anke Holler, University of Göttingen
Graham Katz, Georgetown University
Ralf Klabunde, Ruhr-University Bochum
Valia Kordoni, DFKI GmbH and Saarland University
Ivana Kruijff-Korbayova, Saarland University
Katja Markert, University of Leeds
Costanza Navarretta, University of Copenhagen
Marta Recasens, University of Barcelona
Arndt Riester, University of Stuttgart
Julia Ritz, University of Potsdam
Antje Rossdeutscher, University of Stuttgart
Björn Rothstein, Ruhr-University Bochum
Josef Ruppenhofer, Saarland University
David Schlangen, University of Potsdam
Caroline Sporleder, Saarland University
Manfred Stede, University of Potsdam
Yannick Versley, Tübingen University
Bonnie Webber, University of Edinburgh

Semantics and Pragmatics of Indefinites: Methodology for a Synchronic and Diachronic Corpus Study

Ana Aguilar-Guevara[⋆], *Maria Aloni*[†], *Angelika Port*[†], *Radek Šimík*[‡], *Machteld de Vos*[⋆⋆], *Hedde Zeijlstra*[†]
[⋆]Utrecht University, [†]University of Amsterdam, [‡]University of Potsdam, [⋆⋆]Cambridge University

Abstract
The article discusses the methodology adopted for a cross-linguistic synchronic and diachronic corpus study on indefinites. The study covered five indefinite expressions, each in a different language. The main goal of the study was to verify the distribution of these indefinites synchronically and to attest their historical development. The methodology we used is a form of functional labeling which combines both context (syntax) and meaning (semantics) using as a starting point Haspelmath's (1997) functional map. In the article we identify Haspelmath's functions with logico-semantic interpretations and propose a binary branching decision tree assigning each instance of an indefinite exactly one function in the map.

1 Theoretical Background

It is well known that the use of expressions with existential meaning (e.g. plain indefinites like English *somebody*, or Czech *někdo*) can give rise to different pragmatic effects. Although the semantic representation of *somebody* in (1) and (2) is identical, (1) comes along with a **free choice implicature** (each individual is a permissible option) and (2) with an **ignorance implicature** (the speaker does not know who called):

(1) You can invite somebody.

(2) Somebody called.

From a typological perspective, many languages have developed specialized forms for such enriched meanings, such as **free choice indefinites**[1]: Spanish *cualquier*-series, Czech *koli*-series, Dutch *dan ook*-series, ..., and as **epistemic indefinites**[2]: Russian *to*-series, Czech *si*-series, German *irgend*-series, Spanish *algun*-series, ...

Following Grice's seminal work, the main hypothesis that motivates the present research is that these different indefinite forms have emerged as result of a process of conventionalization (or fossilization) of an originally pragmatic inference.

In languages with Epistemic Indefinite (EI) forms, inference (3c), pragmatic in origin, has been integrated into the semantic content of sentences like (4a).

[1]E.g. Dayal (1998), Giannakidou (2001), Menéndez-Benito (2010).
[2]E.g. Kratzer and Shimoyama (2002), Jayez and Tovena (2006), Alonso-Ovalle and Menéndez-Benito (2010) .

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena", pages 1-16, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

(3) *Plain indefinite (German)*

a. ***Jemand*** *hat angerufen.*
somebody has called
b. Conventional meaning: Somebody called
c. Ignorance implicature: The speaker does not know who

(4) *EI pronoun (German 'irgendjemand')*

a. ***Irgendjemand*** *hat angerufen.*
somebody:UNKNOWN has called
b. Conventional meaning: Somebody called and the speaker does not know who

In languages with distinctive Free Choice (FC) forms, inference (5c) pragmatic in origin, has been integrated into the semantic content of sentences like (6a).

(5) *Plain indefinite (Spanish)*

a. *Puedes traer* ***un*** *libro.*
can:2SG bring:INF a book
b. Conventional meaning: You can bring me a book
c. Free choice implicature: Each book is a possible option

(6) *FC determiner (Spanish 'cualquier')*

a. *Puedes traer* ***cualquier*** *libro.*
can:2SG bring:INF any book
b. Conventional meaning: You can bring me a book and each book is a possible option

In this project, cross-linguistic synchronic and diachronic studies have been combined in order to substantiate this hypothesis. The synchronic studies intend to determine what has been fossilized, the diachronic studies how this has happened.

In the synchronic research we studied the following indefinite forms: German EI *irgendein*, Czech FC *kterýkoli*, Italian FC *(uno) qualunque*, Spanish FC *cualquiera* and Dutch FC *wie dan ook*. The main goal of this research was to understand which part of the meaning of the indefinite form is fossilized and to develop some hypotheses on how it might had happened diachronically. In the diachronic corpus research we studied the historical development of the last two indefinite forms: Spanish *cualquiera* and Dutch *wie dan ook*.

In this article we will focus on the methodology developed for these corpus studies, and report on parts of the diachronic research as an illustration of our results.

2 Corpus study: diagnostics and methodology

In the synchronic and diachronic studies we have classified randomly selected occurrences of each indefinite according to a number of categories. The annotation was carried out by five annotators (one per language) who met regularly to compare their results and share their experience with the annotation instructions.[3] The starting point

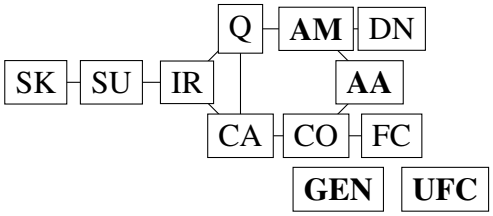
[3] An assessment of the methodology by measuring inter-annotator agreement with the *kappa* coefficient has been carried out in January 2011. Five annotators coded 100 randomly chosen examples from the British National Corpus. Each example contained one marked occurrence of *some* (20 examples) or *any* (80 examples). The average kappa score obtained was 0.52, with a standard deviation of 0.069. We performed a second calculation where the disagreements among the three negative labels (AA, AM and DN) and among the two specific labels (SK and SU) were not taken into account (had a weight of 0), and where the disagreements between the specific functions and IR were considered half correct (had a weight of 0.5). This yielded a kappa score of 0.69, with a standard deviation of 0.106 (for details see van Cranenburgh et al. 2011).

for the identification of the relevant categories was Haspelmath's functional map. In this section, we introduce our extended version of Haspelmath's map and provide an explicit set of logico-semantic criteria, according to which indefinites are assigned functions on the map.

2.1 Haspelmath's semantic map

Haspelmath's (1997) typological survey identified 9 main functions for indefinite forms organized in an implicational map. We will assume the following extended version of Haspelmath's map motivated by a more detailed NPI/FC classification (Aguilar-Guevara et al. 2010). The newly introduced functions are in boldface in the following illustrations:

(7) *An extended version of Haspelmath's map*



(8) *Functions on the map*

	Abbr	Label	Example
a.	SK	specific known	*Somebody* called. Guess who?
b.	SU	specific unknown	I heard *something*, but I couldn't tell what it was.
c.	IR	irrealis	You must try *somewhere* else.
d.	Q	question	Did *anybody* tell you anything about it?
e.	CA	conditional antecedent	If you see *anybody*, tell me immediately.
f.	CO	comparative	John is taller than *anybody*.
g.	DN	direct negation	John didn't see *anybody*.
h.	**AM**	anti-morphic	I don't think that *anybody* knows the answer.
i.	**AA**	anti-additive	The bank avoided taking *any* decision.
j.	FC	free choice	You may kiss *anybody*.
k.	**UFC**	universal free choice	John kissed *any* woman with red hair.
l.	**GEN**	generic	*Any* dog has four legs.

In order for an indefinite to qualify for a function, it must (i) be grammatical in the context the function specifies; and (ii) have the semantics that the function specifies. For example, *any* does not exhibit the specific functions SK/SU because it is ungrammatical in episodic sentences, cf. (9a); and *some* does not exhibit the comparative function CO because it does not have a universal meaning specified by CO, cf. (9b).

(9) a. He went somewhere /# anywhere else.
b. Berlin is bigger than any /# some Czech city.
'For all Czech cities it holds that Berlin is bigger than they are.'

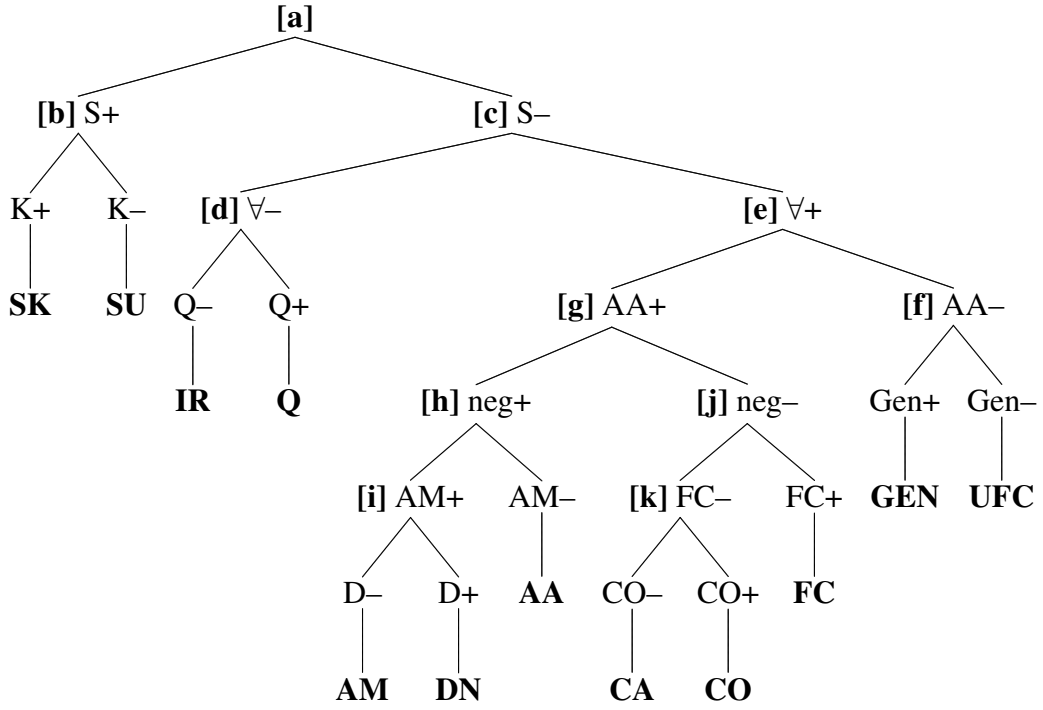
Epistemic indefinites are indefinites that exhibit the SU function, but not the SK function. Free choice indefinites are indefinites exhibiting the FC function.

Haspelmath proposes that an indefinite will always express a set of functions that are contiguous on the map (where two functions are contiguous iff they are connected by a line).[4] One prediction is that items which acquire new functions will develop first those functions that are contiguous to the original function.

2.2 Methodology for semantic annotation

In this section we introduce a set of tests which we used to assign exactly one function to each instance of the examined indefinites. These tests and the order in which they were applied are schematized in the following decision tree.

(10) *Decision tree*



For each node in the decision tree we give now the corresponding test, and, as an illustration, we apply it to the sentences we have used in (8) to exemplify our functional labels. Our first test is test **(a)** used to distinguish specific from non-specific uses of indefinites.

(a) Test for specificity [S+/–]:

Sentence (S): . . . indefinite$_i$ . . . **Possible Continuation (PC)**: . . . pronoun$_i$ . . . [S+]

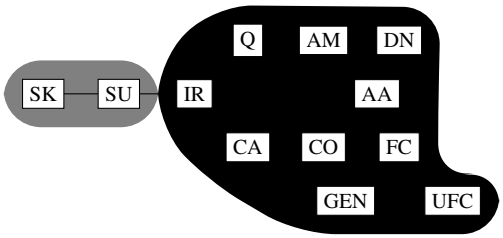
Examples:

a. *Somebody*$_i$ called. She$_i$ wanted a new appointment. [S+]
b. I heard *something*$_i$. It$_i$ was very loud. [S+]
c. You must try *somewhere*$_i$ else. # It$_i$ is a very nice place. [S–]
d. Did *anybody*$_i$ tell you anything about it? # He$_i$ is a real chatterbox. [S–]
e. If you see *anybody*$_i$, tell me immediately. # He$_i$ is a nice guy. [S–]

[4] The precise placement on the map (i.e. connecting lines determining function contiguity) of GEN and UFC is still a matter of investigation.

f. John is taller than *anybody*$_i$. # He$_i$ is short. [S–]
g. John didn't see *anybody*$_i$. # He$_i$ was very tall. [S–]
h. I don't think that *anybody*$_i$ knows the answer. # He$_i$ did not even try. [S–]
i. The bank avoided taking *any* decision$_i$. # It$_i$ was difficult. [S–]
j. You may kiss *anybody*$_i$. # She$_i$ is beautiful. [S–]
k. John kissed *any* woman$_i$ with red hair. # She$_i$ is Italian. [S–]
l. *Any* dog$_i$ has four legs. # It$_i$ is very cute. [S–]

The application of test **(a)** splits our map into a specific area (in grey) and a non-specific area (in black).



Within the specific area we apply test **(b)** to distinguish the specific known from the specific unknown function.

(b) Test for known [K+/–]: **S:**. . . indefinite . . . . **PC:** Guess who/what? [K+]

Examples:

a. *Somebody* called. Guess who? [K+] ↦ [**SK**]
b. I heard *something*, but I couldn't tell what it was. # Guess what? [K–] ↦ [**SU**]

Within the non-specific area we apply test **(c)** to distinguish between wide-scope universal meaning and genuinely existential meaning:

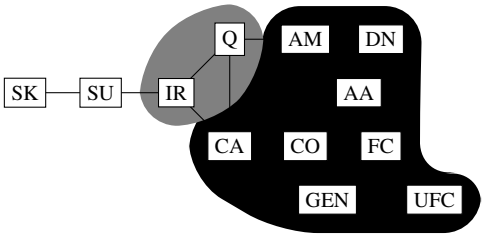
(c) Test for universal meaning [∀+/–]:

. . . **Op** (. . . indefinite . . .) . . . ⇒ . . . ∀x (**Op**. . . x . . .) . . .

Examples:

a. You must try *somewhere* else $\not\Rightarrow$ for every place x: you must try x [∀–]
b. Did *anybody* tell you anything about it? $\not\Rightarrow$ for every x: did x tell you about it? [∀–]
c. If you see *anybody*, tell me immediately ⇒ for every x: if you see x, tell me immed. [∀+]
d. John is taller than *anybody* ⇒ for every x: John is taller than x [∀+]
e. I didn't see *anybody* ⇒ for every x: I didn't see x [∀+]
f. I don't think that *anybody* knows the answer ⇒ for every x: I don't think that x knows the answer [∀+]
g. The bank avoided taking *any* decision ⇒ for every decision x: the bank avoided taking x [∀+]
h. You may kiss *anybody* ⇒ for every x: you may kiss x [∀+]
i. John kissed *any* woman with red hair ⇒ for every woman x with red hair: John kissed x [∀+]
j. *Any* dog has four legs ⇒ for every dog x (with exceptions?): x has four legs [∀+]

The application of test **(c)** splits the non-specific area into an existential area (in grey) and a wide-scope universal area (in black).



Within the existential area we distinguish polar questions from irrealis non-specific constructions via step **(d)**.

(d) Polar question [Q+]

Examples:

a. You must try *somewhere* else. [Q–] ↦ [**IR**]
b. Did you see *anybody*? [Q+] ↦ [**Q**]

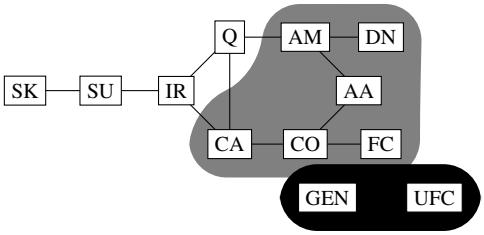
Within the wide-scope universal area we apply test **(e)** to distinguish anti-additive contexts from non anti-additive ones.

(e) Test for anti-additivity [AA+/–]: $\mathbf{Op}(a \vee b) \Rightarrow \mathbf{Op}(a) \wedge \mathbf{Op}(b)$ [AA+]

Examples:

a. If you see *anybody*, you should tell me immediately. [If you see John or Mary, you should tell me immediately ⇒ If you see John, you should tell me immediately and if you see Maria, you should tell me immediately] [AA+]
b. John is taller than *anybody*. [John is taller than Lee or Mary ⇒ John is taller than Lee and John is taller than Mary] [AA+]
c. John didn't see *anybody*. [John didn't see Lee or Mary ⇒ John didn't see Lee and John didn't see Mary] [AA+]
d. I don't think that *anybody* knows the answer. [I don't think that Mary or Lee know the answer ⇒ I don't think that Mary knows the answer and I don't think that Lee knows the answer] [AA+]
e. The bank avoided taking *any* decision. [The bank avoided taking decision A or decision B ⇒ The bank avoided taking decision A and the bank avoided taking decision B] [AA+]
f. You may kiss *anybody*. [You may kiss John or Mary ⇒ you may kiss John and you may kiss Mary] [AA+]
g. John kissed *any* woman with red hair. [John kissed Lee or Bea ⇏ John kissed Lee and John kissed Bea] [AA–]
h. *Any* dog has four legs. [Fido or Bobby has four legs ⇏ Fido has four legs and Bobby has four legs] [AA–]

The application of test **(e)** splits the universal area into an anti-additive area (in grey) and a non anti-additive area (in black).



Within the non anti-additive area we apply test **(f)** to distinguish generic from universal free choice readings.

(f) Test for genericity [Gen+/–]: . . . indefinite . . . ≡ . . . plain generic indef. . . . [Gen+]

Examples:

a. John kissed *any* woman with red hair ≢ John kissed a woman with red hair [Gen–] ↦ [**UFC**]

b. *Any* dog has four legs ≡ A dog has four legs [Gen+] ↦ [**GEN**]

Within the anti-additive area we apply test **(g)** to distinguish negative contexts from non negative ones.

(g) Test for negative meaning [Neg+/–]: **Op**$(a \vee \neg a)$ is inconsistent [Neg+]

Examples:

a. John didn't see *anybody*. [John didn't stay or go ↦ inconsistent] [Neg+]

b. I don't think that *anybody* knows the answer. [I don't think that the door is open or closed ↦ inconsistent] [Neg+]

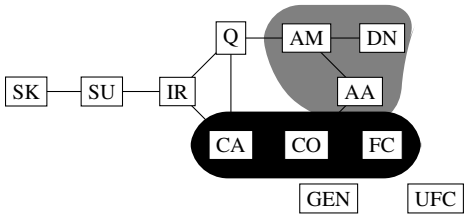
c. The bank avoided taking *any* decision. [The bank avoided being open or closed] ↦ inconsistent] [Neg+]

d. You may kiss *anybody*. [You may stay or go ↦ not inconsistent] [Neg–]

e. If you see *anybody*, you should tell me. [If you stay or go, you should tell me ↦ not inconsistent] [Neg–]

f. John is taller than *anybody*. [John is taller than somebody or nobody ↦ not inconsistent] [Neg–]

The application of test **(g)** splits the anti-additive area into a negative area (in grey) and a non-negative area (in black).



Within the negative area we apply test **(h)** to distinguish anti-multiplicative contexts from plain negative ones.

(h) Test for anti-multiplicativity: **Op**$(a)\vee$**Op**(b) ≡ **Op**$(a \wedge b)$

Examples:

a. John didn't see *anybody*. [John didn't see Mary or John didn't see Sue ≡ John didn't see (Mary and Sue)] [AM+]

b. I don't think that *anybody* knows the answer. [I don't think that Lee knows the answer or I don't think that Mary knows the answer ≡ I don't think that (Lee and Mary) know the answer] [AM+]

c. The bank avoided taking *any* decision. [The bank avoided taking decision A or the bank avoided taking decision B ≢ The bank avoided taking (decision A and decision B)] [AM–] ↦ **[AA]**

Within the anti-multiplicative area we check if the relevant operator is clausal negation.

(i) **Op** is clausal negation [D+]

Examples:

a. John didn't see *anybody*. [D+] $\mapsto$ **[DN]**

b. I don't think that *anybody* knows the answer. [D–] $\mapsto$ **[AM]**

Within the anti-additive non negative area we apply test **(j)** to distinguish free choice contexts.

(j) Test for free choice [FC+/–]: **Op**$(a \vee \neg a)$ is informative [FC+]

Examples:

a. If you see *anybody*, you should tell me. [If you stay or go, you should tell me $\mapsto$ antecedent is not informative] [FC–]

b. John is taller than *anybody*. [John is taller than somebody or nobody $\mapsto$ not informative] [FC–]

c. You may kiss *anybody*. [You may stay or go $\mapsto$ informative] [FC+] $\mapsto$ **[FC]**

Within the non free choice contexts we distinguish the comparative constructions from the others.

(k) Comparative construction [CO+]

Examples:

a. If you see *anybody*, tell me immediately. [CO–] $\mapsto$ **[CA]**

b. John is taller than *anybody*. [CO+] $\mapsto$ **[CO]**

Further applications of the tests Consider now the following ambiguous example from Horn (2005:183):

(11) If she can solve *any* problem, she'll get a prize.

a. ('existential') If there is any problem she can solve, ...

b. ('universal') If she can solve every problem, ...

When applying our decision procedure to this example, at node **(c)** (the test for universal reading) we have to decide on what operator counts as the relevant **Op**. We have two candidates here: the conditional construction or the possibility modal *can*. In the first case (corresponding to the existential reading in (11a)) our terminal node will be **CA**, as illustrated in (12). In the second case, (corresponding to the universal reading in (11b)) our terminal node will be **FC**, as illustrated in (13):

(12) a. If she can solve *any*$_i$ problem, she'll get a prize. # It$_i$ is a very difficult question. [S–]

b. If she can solve *any* problem, she'll get a prize. $\Rightarrow$ For every problem x: (if she can solve x, then she'll get a prize) [$\forall$+]

c. If she solves problem A or problem B, she'll get a prize. $\Rightarrow$ If she solves problem A, she'll get a prize and if she solves problem B, she'll get a prize. [AA+]

d. If she solves or doesn't solve a problem, she'll get a prize $\mapsto$ antecedent is not inconsistent [Neg–]

e. If she solves or doesn't solve a problem, she'll get a prize ↦ antecedent is not informative [FC–]

f. If she can solve *any* problem, she'll get a prize. [CO–] ↦ [**CA**]

(13) a. If she can solve *any*$_i$ problem, she'll get a prize. # It$_i$ is a very difficult question. [S–]

b. If she can solve *any* problem, she'll get a prize ⇒ If (for every problem x: she can solve x), then she'll get a prize [∀+]

c. She can solve problem A or problem B ⇒ She can solve problem A and she can solve problem B [AA+]

d. She can solve a problem or not ↦ not inconsistent [Neg–]

e. She can solve a problem or not ↦ informative [FC+] ↦ [**FC**]

In ambiguous cases like this one, if the context did not disambiguate the intended reading, the sentences were annotated with both possible functions. To keep the randomly chosen occurrences stable the readings were counted as 0.5.

While these tests proved useful for many cases, there were examples for which our decision tree was inconclusive, and we conclude the section by discussing one of these cases. Consider the following example from Horn (2005), (see also Vlachou 2007):

(14) I do not want to go to bed with just *anyone* anymore. I have to be attracted to them sexually.

Applying our tests for specific and for universal reading leads us to place this sentence in the non-specific existential area in our map. This area contains only two functions: Q and IR. Neither of these functions, however, are appropriate for this occurrence since, to quote Horn '*any* appears here in its free choice incarnation' (Horn 2005:185).

(15) a. I do not want to go to bed with just *anyone*$_i$ anymore. # He$_i$ is very handsome. [S–]

b. I do not want to go to bed with just *anyone* anymore. [$\not\Rightarrow$ for every x: I don't want to go to bed with x] [∀–]

c. I do not want to go to bed with just *anyone* anymore. [Q–], but not [IR] either.

To cover these cases we decided to introduce a new function, the indiscriminacy function IND. During annotation we have also introduced other off-map functions to label uses which were not strictly indefinite. One example is the *no-matter* function of which we give here an illustration in Czech:

(16) *A u jsme v kterkoli zemi, vude nachzme slun lidi.*
let already be:1PL in any country everywhere find:1PL polite people
'No matter in which country you are, you can find polite people everywhere.'

In other cases where our decision tree was inconclusive, we left the issue open, and labeled the occurrence as unclear.

3 Some findings

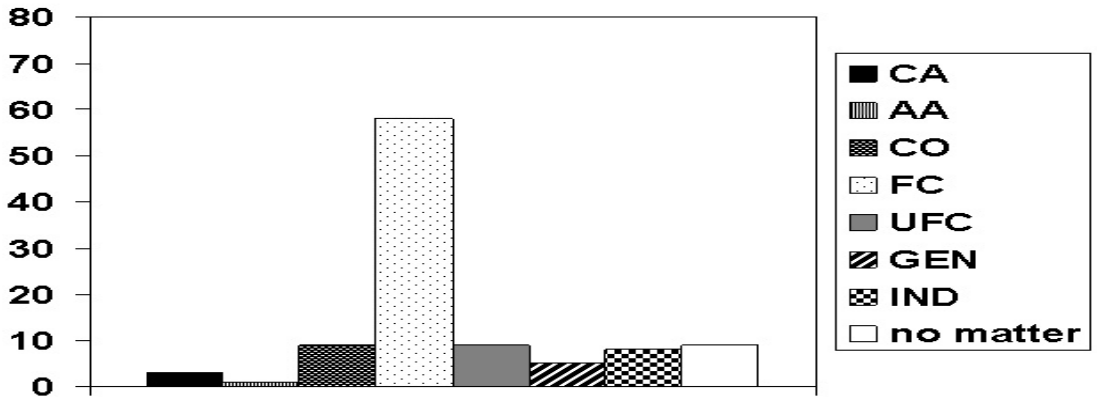
As an illustration of the results of the corpus studies we present the synchronic and the diachronic data of Spanish *cualquiera* and Dutch *wie dan ook*, two constructions that share the property of employing wh-morphology to express free choice meanings.

3.1 Spanish *cualquiera*

For the study of this item, we used *El Corpus del Español* created by Mark Davies. We randomly selected 100 occurrences of *cualquiera* from four sections, namely 1200s (7.9 millions of words), 1500s (19.7 millions of words), 1700s (11.5 millions of words), and 1900s (22.8 millions of words), which represent the four periods in which the history of Spanish has traditionally been divided (cf. Lapesa 1964). We used as a query the sequence *ualq*, which yielded all sorts of spelling variants of the item plus only ten instances of completely unrelated words, which were excluded.

Cualquiera (pronoun), or *cualquier* (determiner), composed of *cual* ('which/who') plus *quier(a)* ('want:3.PRES.SUBJ') has the following distribution in current Spanish:

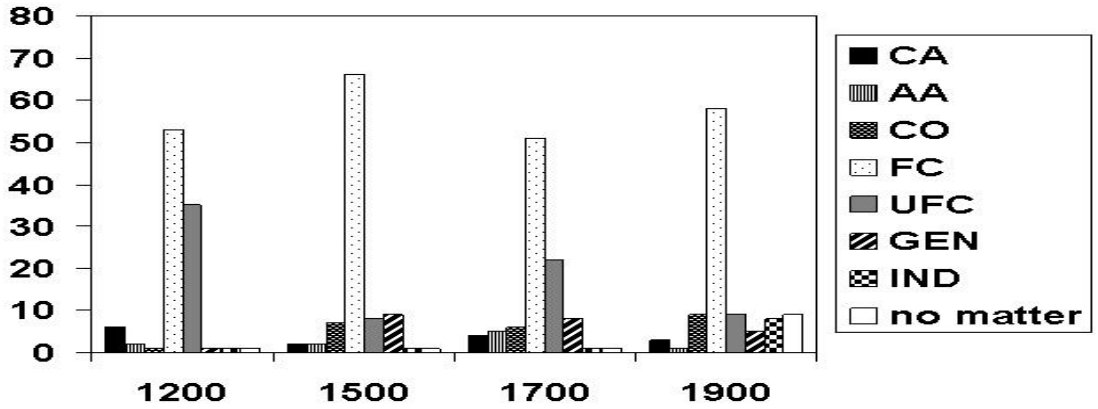
(17) Functions covered by *cualquiera* in current Spanish



This distribution, just like those of the other indefinites discussed in Aguilar-Guevara et al. (2010), confirms Haspelmath's prediction that an indefinite always covers functions that are contiguous in the map.

Let us now discuss the historical development of *cualquiera*. This construction has been claimed to have emerged in Spanish as result of a grammaticalization process through which free relative clauses were reanalyzed as indefinite noun phrases (cf. Company-Company and Pozas-Loyo 2009). Presumably, this process has occurred in early stages of the history of Spanish and in consequence *cualquiera*, as a word, is already recurrently found in the first documentations of Spanish, which date back to the thirteenth century. As discussed in Aguilar-Guevara et al. (2010), the number of instances of *cualquiera* that were documented for each period studied suggest that the use of the construction is already consolidated quite early. The distribution of the functions that *cualquiera* covers throughout these periods points out to a similar conclusion:

(18) Functions covered by *cualquiera* in 1200s, 1500s, 1700s and 1900s



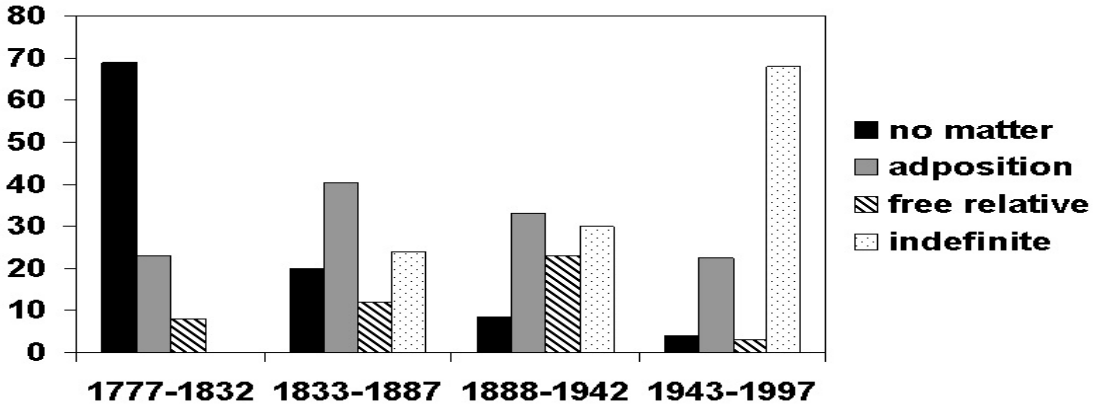
The most noteworthy observation about this distribution is that, generally speaking, it has remained pretty similar throughout the four periods. The FC function is clearly the most dominant since the first period, but some other functions contiguous in the map, namely, CA, CO and AA, as well as the functions UFC and GEN, have some presence as well. Interestingly, the UFC function displays a remarkable decrease as from the 1500s. In Aguilar-Guevara et al. (2010), we tentatively attribute this to the fact that *cualquiera*, as part of its grammaticalization, occurs less and less frequently accompanied by post-nominal modifiers such as restrictive relative clauses and prepositional adjuncts, which typically serve as licensor of free choice items in UFC uses (e.g. *John kissed any woman #(with red hair)*). The last important observation is that two more off-map functions, namely IND and *no-matter*, appear in the 1500s and gain presence by the 1900s. The late emergence of the *no-matter* function will turn particularly interesting in light of the development of the Dutch indefinite *wie dan ook*.

Given the early grammaticalization of *cualquiera* and stable distribution of its functions, we could not really attest much of the process this compound went through in order to behave as it does nowadays. This motivated us to study *wie dan ook*, an indefinite comparable to *cualquiera* in meaning and (partly) in form, but that emerged in Dutch more recently and that even in these days appears to be 'less' grammaticalized than *cualquiera*.

3.2 Dutch *wie dan ook*

The Dutch diachronic study, reported in de Vos (2010), consisted of the analysis of occurrences of *wie dan ook* ('who also then') in written Dutch historical corpora (CD-ROM Middelnederlands (270 texts before 1300), DBNL (4458 texts from 1170-2010)). The first occurrence found is from 1777; the period of this item's existence has therefore been divided into four phases, each covering 55 years of the item's evolution. The outcome shows that *wie dan ook* went through a four-staged process of grammaticalization:

(19) Four stages in grammaticalization of *wie dan ook*



Stage I The first phase in the grammaticalization of *wie dan ook* as an indefinite is formed by three forms of the *no matter*-function. Characteristic of types of *no matter* constructions is that the *wh dan ook* is not part of the main clause yet: they all consist of either a wh-clause and a main clause, or a wh-clause within a main clause, as illustrated as follows:

(20) a. Wie dan ook naar het feest komt; ik zal blij zijn.
'Whoever comes to the party; I will be happy.'

b. [Wie dan ook naar het feest komt]$_i$; hij$_i$ zal blij zijn.
'[Whoever comes to the party]$_i$; he$_i$ will be happy.'

c. Jan, (of) wie dan ook hij mag zijn, zal blij zijn.
'John, (or) whoever he may be, will be happy.'

These forms occur around the same time. Together, they seem particularly frequent in the first phase, forming a significant majority of the total amount of occurrences here, with this relative amount decreasing in the three phases that follow (cf. the black bars in graph (19)).

Stage II In the following stage in the development of *wie dan ook* as an indefinite, *no matter*-constructions are shortened to adpositions, thus getting one step closer to becoming a grammaticalized indefinite. Adpositions have the following form: [. . . , [*wie dan ook*], . . .]. They are shortenings of the *no matter*-function, formed by the ellipsis of the predicate. Although they do not form a separate wh-clause next to or within a main clause anymore, they are still not part of the actual sentence and therefore no real indefinites: they merely modify the noun they are placed after.

(21) Als er iemand$_i$, wie dan ook$_i$, naar het feest komt, zal ik blij zijn.
'If someone, whoever/anyone, comes to the party, I will be happy.'

As the grey bars in (19) show, this adpositional modification with a *wie dan ook* (with ignorance or indifference meaning) is particularly frequent in the second phase in the development of this indefinite.

Stage III The third phase, the *free relative*-stage, shows a further integration of the *wie dan ook*-clause into the sentence, though still not a full integration either. The Free Relative (FR) function, the biggest part of the total amount of occurrences of *wie dan ook* now, forms another spinoff of the *no matter* construction. However, whereas *no matter*-sentences still form combinations of wh-clauses (*wie dan ook* + predicate) and a main clause, the FR-function is more integrated than that, with the "*wie dan ook* + predicate" not forming a separate clause, but an actual part of the main clause, typically the subject. Examples of the FR-function have the following form: [[*wie dan ook* + predicate](,) VP], as illustrated in (22):

(22) Wie dan ook naar het feest komt, zal blij zijn.
'Whoever comes to the party(,) will be happy.'

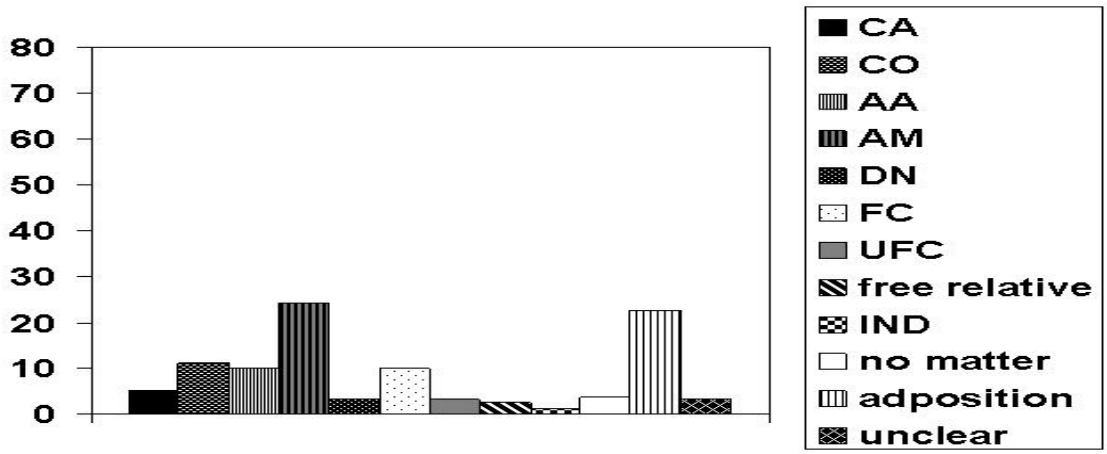
However, these subjects consisting of *wie dan ook* + predicate are often followed by a comma, thereby perhaps indicating that they are still seen as slightly standing outside of the actual sentence. Yet omitting the part starting with *wie dan ook* would give an incomplete thus ungrammatical sentence. This is a specific feature of the third phase; both the *no matter*-clauses and the adpositions can still be left out, of course sometimes causing a change in meaning of the sentence, but never with an incomplete sentence as a result. This shows how integrated a part of the sentence these occurrences of *wie dan ook* already form - although it apparently still feels a bit strange to the contemporary writer. Besides, these forms of *wie dan ook* are not as integrated yet as the plain indefinite will be.

Stage IV In this last stage of the grammaticalization of *wie dan ook*, the word group has finally become an indefinite. Examples of this kind form integrated parts of the sentence, with a plain *wie dan ook*, without any kind of predicate modifying it, being either subject or object: [. . . [*wie dan ook*] . . .].

(23) Je mag wie dan ook uitnodigen voor het feest.
'You may invite anyone to the party.'

Indefinite uses of *wie dan ook* are attested from 1833 onwards, and their number increases in every phase, finally forming a vast majority of the occurrences in the fourth phase, as graph (19) illustrates. Here is the distribution of *wie dan ook* in stage IV:

(24) Functions covered by *wie dan ook* in stage IV (current Dutch)



Summarizing: Overall, what can be concluded is that the process of grammaticalization of *wie dan ook* as an indefinite roughly followed four stages, starting off as a *no-matter* construction in a separate wh-clause, slowly evolving into an adpositional modifier on its own, while also turning into a part of the main clause with predicate, eventually yielding to the true and plain indefinite *wie dan ook* as part of a sentence. Recall that the Spanish study showed a very late emergence of the *no-matter* function for *cualquiera*. This fact, combined with the phases of development of *wie dan ook*, constitutes evidence against unidirectionality in the acquisition of new functions: while the Dutch item was born with the *no-matter* function, the Spanish item starts its development from a free relative into a plain indefinite and only later allows the *no-matter* function to emerge.

Our initial hypothesis was that FC indefinites emerged as the result of a process of conventionalization of an originally pragmatic inference. The envisaged 'conventionalization' is in fact quite difficult to test because conversational implicatures are by definition not overtly expressed. The testing would have to consist in checking for a raising frequency of a conversational implicature of sentences with plain indefinites, then a development of a new morpheme which captures the implicature and then its grammaticalization. Alternatively, the morpheme that had already been used in the plain indefinite would change its function - the implicature would be built in. The latter is not what we observe. Yet, the described development of *wie dan ook* is consistent with the former scenario, with appositive *wie dan ook* as a new form which expresses the original implicature and later gets grammaticalized. More precisely, the grammaticalization path that we are describing for *wie dan ook* could be interpreted as a path from a conversational implicature, via a *conventional* implicature in the sense of Potts (2005)[5] to a conventional meaning (i.e. core / at-issue semantics).

(25) a. Jij mag iemand uitnodigen. (plain indefinite + conversational implicature)
b. Jij mag iemand, wie dan ook (hij mag zijn), uitnodigen. (plain indefinite + conventional implicature)
c. Jij mag wie dan ook uitnodigen (new FC indefinite)

[5] According to Potts (2005), appositives express conventional implicatures, i.e. not at-issue meanings.

To conclude, the emergence of *wie dan ook* as a plain indefinite counts as a classical example of grammaticalization, where the initial periphrastic usage of a wh-clause increased in frequency to such an extent that this usage got reanalyzed as being part of its lexical semantics. Such a process, as is often attested, takes place in a step-wise fashion. The adpositional usage results from the *no matter* usages of *wh*-clauses and can be taken to be the first lexicalization of a FC implicature. However, this adposition brings in new usage effects as well, such as its strong collocational distribution w.r.t. subjects and objects. This, in turn, then causes the next steps of the grammaticalization process: the replacement of DPs by the *wh*-element. Grammaticalization is thus not a big step from a lexical to a functional category (*in casu* from a wh-clause towards an indefinite), but a series of small steps, each possibly being the result of lexicalization of implicatures.

4 Conclusion

We have discussed the methodology adopted for a cross-linguistic synchronic and diachronic corpus study on free choice and epistemic indefinites. The study covered five indefinites in five languages. The main goal of the study was to verify the distribution of these indefinites on an extended version of Haspelmath's (1997) functional map, and to attest their historical development. One of the main conclusions of the synchronic studies was that there is no indefinite that violates the function contiguity. An interesting conclusion of the diachronic research was that the acquisition of new functions is not unidirectional. These studies could not confirm, but neither reject, our initial hypothesis on implicature fossilization.

References

Ana Aguilar-Guevara, Maria Aloni, Angelika Port, Katrin Schulz, and Radek Šimík. Free choice items as fossils. Workshop on Indefiniteness Crosslinguistically (DGfS) Berlin, February 25/26, 2010.

Luis Alonso-Ovalle and Paula Menéndez-Benito. Modal indefinites. *Natural Language Semantics*, 18:1–31, 2010.

Concepción Company-Company and Julia Pozas-Loyo. Los indefinidos compuestos y los pronombres genérico-impersonales *omne* y *uno*. In Concepción Company-Company, editor, *Sintaxis histórica de le lengua española (Segunda parte: La frase nominal)*, pages 1073–1219. Fondo de Cultura Económica-Universidad Nacional Autónoma de México, México City, 2009.

Veneeta Dayal. *Any* as inherently modal. *Linguistics and Philosophy*, 21:433–476, 1998.

Machteld de Vos. Wh dan ook: The synchronic and diachronic study of the grammaticalization of a Dutch indefinite. BA thesis, University of Amsterdam, 2010.

Anastasia Giannakidou. The meaning of free choice. *Linguistics and Philosophy*, 24:659–735, 2001.

Martin Haspelmath. *Indefinite pronouns*. Oxford University Press, Oxford, 1997.

Laurence Horn. Airport '86 revisited: Toward a unified indefinite *any*. In Gregory Carlson and Francis J. Pelletier, editors, *The Partee Effect*, pages 179–205. CSLI, Stanford, 2005.

Jacques Jayez and Lucia Tovena. Epistemic determiners. *Journal of Semantics*, 23:217–250, 2006.

Angelika Kratzer and Junko Shimoyama. Indeterminate pronouns: The view from Japanese. In

Yukio Otsu, editor, *The proceedings of the Third Tokyo Conference on Psycholinguistics*, pages 1–25, Tokyo, 2002.

Rafael Lapesa. *Historia de la lengua española*. Gredos, Madrid, 1964.

Paula Menéndez-Benito. On universal Free Choice items. *Natural Language Semantics*, 18:33–64, 2010.

Christopher Potts. *The Logic of Conventional Implicatures*. Oxford University Press, Oxford, 2005.

Andreas van Cranenburgh, Raquel Fernandez, Katya Garmash, Marta Sznajder, and Maria Velema. Assessing the reliability of an annotation scheme for indefinites. Technical Report, ILLC, University of Amsterdam, 2011.

Evangelia Vlachou. *Free Choice in and out of Context: The semantics and distribution of French, Greek and English Free Choice Items*. PhD thesis, LOT dissertation series 156, Utrecht, 2007.

Syntax-Centered and Semantics-Centered Views of Discourse. Can They be Reconciled?

Matthias Buch-Kromann, Daniel Hardt, and Iørn Korzen
Copenhagen Business School

Abstract

In this paper, we argue that there are two seemingly incompatible perceptions of discourse structure: a semantics-centered view and a syntax-centered view. In the semantics-based view, discourse structure is viewed as a structure that identifies the most important portions of the text and describes how they combine semantically. In the syntax-based view, discourse structure is viewed as an extension of syntax to the discourse level, which essentially links the syntactic trees for the individual sentences into one big tree structure. We will argue that these differences in perception may explain some of the central disagreements in the literature about the nature of discourse structure, in particular whether discourse structure is best viewed as a tree or a general graph. However, the two views are not as incompatible as they may seem at first sight, since the semantic discourse structure can be reinterpreted as a functor-argument structure that is derived from the syntactic tree structure. We describe the ramifications of the two views for the analysis of discourse markers, which are the focus of the discourse annotation in the Penn Discourse Treebank, and show how the syntax-based view can maintain a tree structure even for examples that seem to exhibit non-tree like properties in a semantics-based view.

1 Introduction

Most research on discourse structure builds on the premise that coherent texts have an associated internal structure that places constraints on how the meaning of the whole text is computed from the meanings of its individual clauses and sentences, and how the individual clauses and sentences are presented in the linear order. When texts appear coherent and easily comprehensible to readers, it is because they have a well-formed discourse structure that respects the listeners' conventions about linear order and sensible semantic and pragmatic interpretation, whereas texts that lack this property are perceived as being incoherent and difficult to comprehend. Discourse structure is mostly viewed as a tree structure, or at least a very tree-like structure, where the most important portions of the text are assumed to be located at or near the top of the tree, and the deepest parts of the tree are supposed to encode supplementary information that is less central to the writer's purposes and can be more easily excluded from a summary of the text. The individual branches in the discourse tree define discourse units which are supposed to form coherent textual units that can be interpreted in isolation, a property that can be used in reverse to identify the discourse units in a text. To varying degrees, this general framework forms the theoretical basis for discourse theories like Rhetorical Structure Theory (Mann and Thompson, 1987), the Linguistic Discourse Model (Polanyi, 1988), and Segmented Discourse Representation Theory (Lascarides and Asher, 2007), and for discourse treebanks like the English RST treebank (Carlson et al., 2001), the Discourse Graphbank (Wolf and Gibson, 2005), the Penn Discourse Treebank (Prasad et al., 2008) and related discourse treebanks (Mladová et al., 2008; Aktaş et

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 17-30, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

al., 2010), the Potsdam Commentary Corpus (Stede, 2008), and the Copenhagen Dependency Treebanks (Buch-Kromann and Korzen, 2010).

This mainstream view of discourse structure is to a very large degree inspired by the success with which syntactic theory has managed to account for intra-sentential structure. In mainstream syntax, the structure of a sentence is modelled by means of a tree augmented with additional structure which may be used to handle semantics or deal with non-canonical word order and secondary dependencies[1] (e.g., in topicalizations, control constructions, and relative clauses); this is true for a wide range of syntactic theories, including Head-Driven Phrase Structure Grammar[2] (Pollard and Sag, 1994), Lexical-Functional Grammar (Dalrymple et al., 1994), Government and Binding Theory (Chomsky, 1981/1993), Combinatory Categorial Grammar (Steedman, 2000), Tree-Adjoining Grammar (Joshi and Schabes, 1997), and different versions of dependency grammar (Hudson, 2010; Mel'čuk, 1988; Sgall et al., 1986; Duchier, 2001; Buch-Kromann, 2009; and many others). It is therefore tempting to try to reuse these mechanisms for the analysis of discourse, which is what most theories of discourse have sought to do (with Wolf and Gibson (2005) as the clearest exception). In syntax, there seems to be agreement about the general mechanisms needed to account for syntactic structure, although the specific implementational details vary greatly between the frameworks; but in discourse, there is a much lower level of agreement about the detailed theoretical interpretation and function of discourse structure and its relationship to syntax and discourse semantics.

In this paper, we seek to clarify some of these interpretational problems. By drawing on the insights and mechanisms from syntax and its relationship to sentential semantics, we hope to shed light on ways in which these insights may be carried over to our understanding of discourse. The paper is structured as follows. In section 2, we describe the blurry syntax-discourse boundary and the implications for the relationship between syntactic structure and discourse structure. In section 3, we describe the syntactic distinction between constituent structure and functor-argument structure, and argue in section 4 that this distinction is relevant for discourse as well. In section 5, we discuss the implications for the analysis of discourse connectives. In section 6, we argue that attribution is a particularly hard problem for a tree-based analysis of discourse, but that the problem can be resolved by either a more careful semantic analysis or a small extension of the compositional semantics. In section 7, we revisit some of the counter-examples that have been used to argue against a tree-based view of discourse structure. In Section 8, we identify some of the outstanding problems in a syntax-based view of discourse. In section 9, we describe how these insights have informed the syntax-based discourse annotation in the Copenhagen Dependency Treebanks. Our conclusions are presented in section 10.

1 By a secondary dependency we mean the phenomenon that a single phrase may sometimes function as a complement or adjunct in several phrases simultaneously, e.g., in control constructions where the control verb licenses a subject to function as a secondary subject of the controlled verb, or in relatives where the relativized noun functions as a secondary complement or adjunct within the relative clause, in addition to its external syntactic role.

2 Although HPSG analyses are directed acyclic graphs, many of the HPSG features encode trees, e.g., the DTRS feature.

2 The blurry syntax-discourse boundary and the interface problem

As noted by Carlson and Marcu (2001), the boundary between syntax and discourse is rather fuzzy, and the same meaning can be expressed in a continuum of ways that range from clear discourse constructions ("He laughed. That annoyed me.") to clear syntactic constructions ("His laugh annoyed me."). Discourse and syntax may also interact in complicated ways. For example, long discourse units that span several sentences may function as objects of attribution verbs in direct or indirect speech, or as parenthetical remarks embedded within an otherwise normal sentence, and Wolf and Gibson (2005) and Buch-Kromann and Korzen (2010) provide examples where a complex discourse unit elaborates on a preceding NP. This raises obvious questions about how syntactic structure, which is well understood, relates to discourse structure. Since most discourse frameworks take the clause as their minimal discourse unit, there is some overlap where we can compare the intra-sentential discourse structure with the corresponding syntactic structure. When these structures differ, we must ask why they differ and how they interface with each other, given that they serve the same purpose of determining the compositional semantics and controlling the linear order, but at different linguistic levels.

It is important to note that at the intra-sentential level, discourse frameworks frequently provide structural analyses that differ from the corresponding syntactic structure, even when there is near-universal agreement about the syntactic analysis across syntactic frameworks. For example, in attributions like "The children said that they liked ice cream", the subordinate clause "that they liked ice cream" is universally analyzed as the syntactic complement of the main clause "The children said...", whereas discourse frameworks as implemented in the RST Treebank and GraphBank reverse the direction by analyzing the attribution clause as a subordinate of the attributed clause. Similarly, in discourses like "On the one hand, *X*. On the other hand, *Y*.", the two discourse adverbials "on the one hand" and "on the other hand" are universally analyzed in syntactic theories (including Lexicalized Tree-Adjoining Grammar (LTAG)) as adverbials that modify *X* and *Y*, respectively; but in D-LTAG and PDTB (Webber, 2004; Forbes-Riley et al., 2006), the two adverbials are analyzed as a single lexical item that takes *X* and *Y* as its arguments, reversing the direction of the subordination compared to syntax. In discourse (1) below, the mainstream syntactic analysis is a tree structure where "then" is analyzed as an adverbial and "when" as a subordinating conjunction; in the PDTB analysis, "then" is analyzed as the lexical anchor of an elementary tree that takes the italicized and boldfaced clauses as its argument, resulting in a completely different structure that may even be a non-tree if "when" is assumed to represent a discourse connective as well.

(1) *In an invention that drives Verdi purists bananas, Violetta lies dying in bed during the prelude, rising deliriously* **when** then **she remembers the great parties she used to throw**. (PDTB manual, example (36))

These differences between syntactic structure and the D-LTAG conception of discourse structure is not a problem in itself, since D-LTAG explicitly seeks to model the semantic rather than syntactic structure of discourse. But it does make it harder to reconcile PDTB's semantic conception of primary linguistic structure with the purely syntactic conception found in syntax. In the remainder of this paper, we will argue that we can reconcile the two views within a syntax-centered conception of discourse, by reframing the semantics-centered D-LTAG and PDTB conception of discourse structure as the implicit functor-argument structure associated with a single unified syntax-discourse tree structure for the entire discourse, whose elementary segments represent individual lexical items (typically words).

3 Syntax: syntactic structure vs. functor-argument structure

Syntactic theories almost universally represent syntactic structure as a tree, or a more general graph that has a primary tree as its explicit or implicit backbone, which encodes the syntactic relationships between the constituents in the sentence and constrains their linear order. The main functions of the primary tree are to control the word order and to provide an interface to semantics. Most formal semantic theories assume that phrases are assigned meanings according to the principle of compositionality, which states that the meaning of a phrase is computed as a function of the meanings of its parts (possibly supplemented with some kind of representation of the context in a dynamic semantics). We will follow Dowty (1992) and the approach taken in many linguistic theories, including HPSG, by assuming that complements are lexically selected by their governor and function as semantic arguments to their governor in the compositional semantics, whereas adjuncts lexically select their governor and function as modifiers to their governor in the compositional semantics. The intuition behind Dowty's proposal is that if we have a phrase *XP* with lexical head *X*, complement phrases $C_1,...,C_m$, and adjunct phrases $A_1,...,A_n$ (in increasing scope order), then the meaning [*XP*] associated with the phrase is computed by first applying the functor *h* associated with the lexical head *X* of the phrase to the meanings $[C_1],...,[C_m]$ associated with the complements, and then applying the adjuncts $[A_1],...,[A_n]$ in scope-order, i.e., we define:

$$[XP] = [X + C_1...C_m + A_1...A_n]$$

which is in turn defined recursively by:

$$[X + C_1...C_m] = h([C_1], ..., [C_m])$$
$$[X + C_1...C_m + A_1...A_k] = a_k([A_k])\ ([X + C_1...C_m + A_1...A_{k\text{-}1}]$$

where a_k denotes the functor associated with the adjunct role used to incorporate adjunct A_k into the meaning associated with *XP*. That is, in the syntax, the syntactic head defines the syntactic properties of the entire phrase, but semantically, each adjunct functions as a semantic head, i.e., it acts as a special kind of functor (modifier) that takes *X* with its complements and lower-scoped adjuncts as its argument.

Obviously, the order in which the adjuncts are applied in this meaning computation (the adjunct scope) may affect the meaning we compute for the entire phrase, i.e., two different scopes may (or may not, depending on the circumstances) lead to different meanings. Less obviously, this view of compositional semantics does not necessarily imply a conception of meaning composition as function application: it may be the case that the meaning representation associated with [*XP*] contains the meaning representations $[C_1]$,...,$[C_m]$ and $[A_1]$,...,$[A_n]$ as proper substructures, but the relationship could be more complicated. For example, the meaning composition could be non-monotonic by allowing functors to change or augment substructures in the argument representations (e.g., in the treatment of free subject predicatives that act as adjuncts of the verb, although they really modify the subject from a semantic point of view). Likewise, in a dynamic semantics (cf. Groenendijk and Stokhof, 1991), the meaning composition might imply updates to the hearer's representation of the context; in such a model, expressions like parentheticals could conceivably be modelled as modifiers that affect the context exclusively without affecting the meaning of the phrase that they modify.

The distinction between phrase structure and functor-argument structure has been important in the theoretical development of syntax because it makes it possible to have two structures that serve very different purposes: a surface syntactic structure that essentially controls syntactic constraints on word order, agreement, secondary dependencies in relatives and control constructions3, etc.; and a functor-argument structure that allows for a rich and complex interface to a powerful notion of semantics, while retaining a close and well-defined interface to the syntactic structure via the notion of modifier scope. This realization did not come easily in syntax, as witnessed by the large literature on headedness in syntax (cf. Hudson, 1987; Croft, 1995; Manning, 1995).

We believe that this observation should be of interest to current theories about discourse, which do not currently seem to embody a clear distinction between syntactic and semantic structure. In their annotations, most discourse frameworks seem to lean towards a semantics-centered view where the annotations primarily encode semantic units (corresponding to the intermediate meaning representations in a functor-argument tree) and the relations between them. Today, the field seems to have moved from an initial assumption that a single tree structure may simultaneously explain the semantic interpretation and the syntactic linearization properties of discourse structure (e.g., Mann and Thompson, 1987; Polanyi, 1988; Carlson et al., 2001), to an appreciation that there do exist counter-examples where it seems difficult to find a single tree structure that reconciles these two conflicting requirements (e.g., Wolf and Gibson, 2005; Dinesh et al., 2005; Stede, 2008; Aktaş et al., 2010).

The conception of syntactic structure as a primary tree (possibly augmented with other relations) would have seemed just as untenable in syntax if syntax had been restrained to accounting for both phrase structure and functor-argument structure by means of a single tree. It therefore seems worthwhile asking whether the mechanisms

3 see footnote 1

that appear to have worked so well for syntax could be applied equally successfully to discourse, and what disadvantages, if any, would be associated with a shift from a semantics-centered to a syntax-centered view of discourse.

4 Discourse connectives: heads, conjunctions, markers, or adverbials?

To compare a semantics-centered and a syntax-centered conception of discourse, it is instructive to take a closer look at the analysis of discourse connectives. Discourse connectives form the backbone of the discourse annotations in the Penn Discourse Treebank and the discourse treebanks it has inspired for other languages, and seem crucial in discourse parsing: their presence as simple syntactic clues to the choice of discourse relation probably offers the best chance of getting a hold on a complex linguistic structure which is as ambiguous as it is challenging in terms of its semantic and pragmatic interpretation.

Discourse connectives are typically constructions of the form "*X C Y*", where *X* and Y are clauses and *C* is a discourse connective (such as "because", "since", "when"). Three syntactic analyses and one anaphoric analysis suggest themselves, as summarized in Table 1[4]. The analyses are drawn as dependency trees, i.e., all nodes in the tree represent elementary textual units, and the arrows go from the lexical head of a phrase to the lexical heads of its complement and adjunct phrases, with the relation name written at the arrow tip; the relation name uniquely identifies whether the dependent is a complement or adjunct. Dependency trees can be viewed as being isomorphic to restricted phrase-structure trees where every phrase has a lexical head, but depart from traditional phrase-structure trees in that discontinuous phrases (crossing branches) are allowed.[5]

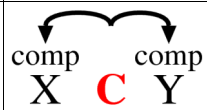
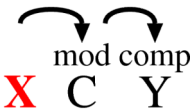
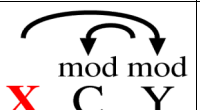
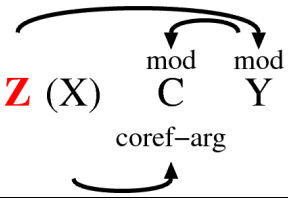
	Head	**Conjunction**	**Marker**	**Anaphoric adverbial**
Syntactic head	C	X	X	Z
Semantic head	C	C	Y	C
Syntax				
Semantics	C'(X',Y')	[C'(Y')] (X')	[Y'(C)] (X')	[[C'(X'')] (Y')] (Z')
CDT example	X said Y	X because Y	X and Y.	Z said X. Then Y.

Table 1. Four analyses of discourse connectives.

In the first analysis (the *head analysis*), the discourse connective *C* is analyzed syntactically as a head that takes *X* and *Y* as its complements; semantically, the meaning *C'* of *C* acts as functor, and the meanings *X'*,*Y'* of *X*,*Y* act as arguments of *C'*.

4 The bottom row of Table 1 provides examples from the Copenhagen Dependency Treebanks (CDT)

5 Each node in the dependency tree corresponds to a possibly discontinuous phrase consisting of the yield of the node in the tree, i.e., the set of all nodes that can be reached by following the dependency arrows in the tree.

In the second analysis (the *conjunction analysis*), *C* is analyzed as a subordinating conjunction that takes *Y* as its complement and modifies *X* as an adjunct; semantically, *C*' computes a meaning *C*'(*Y*') from *Y*', which acts as functor with *X*' as argument. In the third analysis (the *marker analysis*), *C* is analyzed as a marker that modifies *Y*, which in turn modifies *X*; semantically, *Y* selects a composition function *Y*'(*C*) from an inventory of composition functions associated with the head of *Y*, and the semantically vacuous marker *C* merely helps disambiguate the composition function; the composition function then takes the meaning *X*' as its argument. In the final analysis (the *anaphoric adverbial analysis,* initially suggested by Creswell et al. (2002)), the discourse connective *C* retrieves its first argument *X*'' anaphorically, i.e., *C* contains an implicit anaphor *X*'' that provides the first argument in the discourse relation and has *X* as its antecedent. Syntactically, *C* is an adverbial that modifies *Y*, which in turn modifies some other discourse unit *Z* (we do not place any apriori restrictions on *Z*: it might be *X* itself, a unit containing *X*, or a completely different unit). Semantically, *C* is analyzed as a conjunction, i.e., *C*' computes a meaning *C*'(*Y*') from *Y*', which in turn acts as functor with *X*' as argument; the resulting meaning [*C*'(*Y*')](*X*') then acts as a functor which takes *Z*' as its argument. Note that this analysis assumes that two discourse relations are involved: one between *X* and *Y*, and another between *Z* and *Y* (possibly with a completely different connective).

The four analyses are markedly different in terms of their syntactic and semantic headedness, but similar in terms of their semantics, where *X*',*Y*' end up as arguments in all four cases (via a reference *X*'' to *X*' in the anaphoric analysis). If the discourse connective is optional, which is very often the case, the marker analysis has the obvious benefit that there is no need to postulate the presence of an implicit phonetically empty connective: the choice of composition function must then be disambiguated on the basis of semantic and pragmatic clues, rather than overt syntactic clues. This analysis also implies that since discourse markers always modify the satellite, explicit and implicit discourse markers can be used to determine the discourse relation and its direction.

Since the Penn Discourse Treebank only annotates explicit and implicit connectives, with their two arguments, the annotation itself does not specify which of the four syntactic analyses defined above applies to the individual annotations. But from the work on D-LTAG (Forbes-Riley et al., 2006), the theoretical framework that informs the annotation of the Penn Discourse Treebank, it appears that D-LTAG analyzes subordinating conjunctions like “although” as initial trees (essentially a head analysis), coordinating conjunctions like “and” and “but” are analyzed as auxiliary trees (essentially a conjunction analysis, with a phonetically empty connective if the connective is implicit), discourse adverbials like “then” are analyzed as discourse adverbials, and parallel adverbial constructions like “On the one hand, X. On the other hand, Y” are analyzed as initial trees (head analysis). Interestingly, although D-LTAG is based on the syntactic framework LTAG, D-LTAG differs from LTAG in its analysis of subordinating conjunctions and parallel adverbial clauses: D-LTAG uses a head analysis for these constructions, instead of the conjunction analysis an

adjunct analysis used in LTAG and most other syntactic frameworks.[6]

5 Attribution: a difficult case requiring the full power of compositional semantics

As pointed out by Dinesh et al. (2005), attribution is one of the main obstacles for a syntax-centered conception of discourse. Consider their discourse analysis in (2) below:

(2) *The current distribution arrangement ends in March 1990*, although Delmed said **it will continue to provide some supplies of the peritoneal dialysis products to National Medical**, the spokeswoman said. [(12) in Dinesh et al.]

Ignoring the final attribution to the spokeswoman, the discourse is of the form "*X* although Delmed said *Y*". The problem here is that mainstream syntax universally analyzes "Delmed said *Y*" as the complement of "although", but the most sensible reading of (2) is that the discourse relation signalled by "although" holds between *X* and *Y*, rather than between *X* and Delmed's saying event. Carlson et al. (2001) and Wolf and Gibson (2005) try to circumvent this problem by analyzing the attribution as a satellite and the attributed event as the nucleus, but this does not really solve the problem, since the discourse relation may also refer to the attribution event, as demonstrated by (3):

(3) *Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor,* while **opponents argued that the increase will still hurt small business and cost many thousands of jobs**. [(13) in Dinesh et al.]

Dinesh et al. suggest that the problems with attribution could be taken as arguments against a tree-structured discourse, which would undermine a syntax-based view of discourse. We would like to propose two alternative responses – the first accepts the analysis of these examples given by Dinesh et al., while the second proposal relies on a different analysis.

Our first proposal involves the introduction of a more powerful compositional mechanism to address the problem pointed to by Dinesh et al. Given the highly complex compositional semantic mechanisms that are needed in syntax, in any case (e.g., for markers and Pustejovsky-style lexical semantics), we find this is a reasonable response, rather than giving up the idea that discourse structure can be modelled by a syntactic tree.

Specifically, suppose we have a discourse of the form "*X C Y*" where *X* and *Y* may contain a chain of attributions (i.e., *Y* could be of the form "Delmed said *Z*", "Delmed said Ann claimed *Z*", "Delmed said Ann claimed Bob believed *Z*", etc.). Let *c* denote the standard composition function associated with *C*, and suppose π is an operator

6 Since the purpose of D-LTAG is to perform discourse parsing, it is quite possible that this change in analysis is motivated by computational rather than linguistic considerations.

that given an epistemic formula $K_a\varphi$ ("φ is known by agent a") returns φ. In order to handle attributions in the compositional semantics, we only have to assume that instead of letting C have a single composition function c which given arguments X',Y' computes a meaning representation $c(X',Y')$, it has a whole family of composition functions c_{ij} defined by $c_{ij}(X',Y') = c(\pi^i(X'),\pi^j(Y'))$ where i,j cannot exceed the length of the attribution chain in X,Y. When computing the compositional semantics, we then have to disambiguate not only the correct relation associated with C, but also the correct choice of i,j.

This step is not as radical as it may seem at first sight. Many explicit discourse connectives seem to support more than one reading, i.e., they have more than one natural composition function. If we also adopt the marker analysis, we are in principle assuming that any discourse unit can attach to any other discourse unit, choosing a composition function from the full inventory of discourse relations on the basis of contextual clues and optional syntactic clues. In this case, our compositional treatment of attribution essentially just means adding a little more ambiguity to the set of composition functions provided by the inventory of possibly implicit discourse relations.

The compositional account of attribution does not prevent us from making a precise annotation either, since we can disambiguate the correct choice of numbers i,j for a relation R by annotating the relation as "iRj" rather than "R" – this is actually the essence of the annotation scheme for attribution used in the Copenhagen Dependency Treebanks (Buch-Kromann and Korzen, 2010), except that i and j are annotated as sequences of asterisks, rather than as numbers. Attribution is therefore not as big an obstacle to a syntax-centered conception of discourse that it might at first appear to be.

Our second response calls into question the analysis given of example (2) by Dinesh et al. – the key problem is that *although* relates X with Y, rather than relating X with "*Delmed said Y*". The syntax-discourse mismatch is eliminated if it is possible to analyze "*Delmed said Y*" as the second argument of the contrast relation, and we argue that this indeed is the proper analysis here. In fact, it is typical for contrastive relations to arise between conflicting propositions from different sources: in fact that is precisely the situation in example (3), as Dinesh et al. point out. The only difference in (2) is that the first argument is *implicitly* associated with the speaker, while the second argument is explicitly associated with Delmed. In our view, it is quite natural to contrast the two under the assumption that Delmed is credible.

It may well be that there are cases of attribution that require an analysis that reveals a mismatch between syntax and discourse. But in our view, examples (2) and (3) from Dinesh et al. are properly analyzed without any such mismatch. Thus while we are open to the possibility that the more complex compositional mechanism may indeed be necessary, we leave the issue unresolved in this paper.

6 Tree structured discourse: the counter examples from a syntax-centered view

A lot of research in discourse structure has centered on the question whether discourse structure can be viewed as a tree structure or not. Wolf and Gibson (2005)

were among the first to question the suitability of tree structures for discourse, followed by many other researchers, including Dinesh et al. (2005), Lee et al. (2006, 2008), Stede (2008), and Aktaş et al. (2010).

Wolf and Gibson (2005) created a corpus of discourse analyses, without requiring the analyses to be trees, and found that the resulting analyses deviated significantly from trees by including crossing relations and multi-nuclearity. In a syntax-centered conception of discourse, Wolf and Gibson's finding with respect to crossing relations only shows that discourse resembles syntax in this respect, since discontinuous word order phenomena are a key issue in syntactic frameworks, and all sophisticated syntactic theories have a complex set of mechanisms to account for this challenge. Multi-nuclearity is much harder to reconcile with a syntax-centered view of discourse, but here we essentially agree with the counter-criticism voiced by Marcu (2003), who argued that some of the additional relations were really coreference relations, and the remaining counter-examples might be an artefact of their annotation conventions; this view is mostly supported by Knott (2007).

Dinesh et al. (2005) compared the annotations of subordinating conjunctions in the Penn Discourse Treebank (PDTB) with the syntactic annotations in the Penn Treebank (PTB). They found that there were significant differences between the analysis of syntax and discourse, most of which were caused by the treatment of attribution in the PDTB. The problems associated with attribution was addressed in the preceding section, and we believe some of their other counter-examples can be explained by other means: in some cases, the analysis is ambiguous in both the syntactic annotation and the discourse annotation, and the PTB annotators did not choose the same analysis as the PDTB annotators (e.g., their examples (14)-(15)); in other cases, the analysis chosen by PDTB could have been obtained by assuming a particular modifier scope in the syntactic analysis (e.g., their examples (16)-(17)); differences may also be caused by the coarser granularity of the segmentation in the PDTB (e.g., their examples (18)-(19)).

Lee et al. (2006) provide additional examples of complex discourses from the PDTB that violate one or more tree constraints, including examples of independent relations, shared arguments, properly contained arguments, pure crossings, and partially overlapping arguments. We will follow their formatting conventions, using boldface for the arguments of the first connective, and italics for the arguments of the second connective. Their example of shared arguments has the form "**X** but ***Y*** so *Z*", which looks very different from a syntax tree, especially because they seem to draw the functor-argument tree rather than the syntax tree (if it was a syntax tree, they would be using a head analysis). In the mainstream syntactic analysis of this example, "so Z" modifies "Y" and "but Y so Z" modifies "X", with the connectives analyzed as either conjunctions or markers, depending on personal preference (cf. section 4). This would give the functor-argument structure "**X** but ***Y* so *Z***", i.e., we can obtain the same semantic analysis as in the PDTB if we allow the compositional semantics of "but" to strip off "so Z" from "Y so Z" before the composition with "X", a strategy that does not seem to be completely untenable, given the complexity of the compositional semantics in many other respects. We believe that this

mechanism, coupled with the anaphoric discourse adverbial analysis proposed by Forbes-Riley et al. (2006), can explain the examples of properly contained arguments, pure crossings, and partially overlapping arguments given by Lee et al. In their example of independent relations, consisting of two unconnected trees, the second tree could be analyzed as an elaboration of an NP in the first tree. The examples provided by Aktaş et al. (2010) follow essentially the same pattern as in Lee et al., and we believe they can be accounted for by means of the same mechanisms.

Stede (2008) considers a range of criteria that could be used to determine the analysis of nuclearity in a discourse, including the intention of the text (which segment is most central to the writer's purposes), the thematic development of the text (recurrence, repetition, digression meta-discursive element), surface-oriented properties (connectives, other lexical marking, syntactic structure), and specific conventions adopted by the annotation scheme. He argues that these criteria are often conflicting, in particular, that it is possible to find examples where the writer's purposes run against the syntactic subordination. These counter-examples are typically of the form "*X Y Z*", where *X* and *Y* are related by a multi-nuclear relation, i.e., either of them could function as the nucleus, and *Z* can be manipulated so that it is a satellite of either *X* or *Y*. Stede's argument is that in a discourse structure based on trees where crossing relations are disallowed, we will be forced to select different analyses of the relationship between *X* and *Y*, i.e., the nucleus of "*X Y*" necessarily coincides with the nucleus for *Z*. However, these examples could just as well be taken as evidence for crossing relations, which would not be problematic in a syntax-based conception of discourse. In other cases where Stede departs from the syntactic analysis in his discourse analysis, he does so because he sees the syntactic structure as peripheral to Mann and Thompson's characterization of the nucleus as being "more central to the writer's purposes", i.e., he essentially reverses the syntactic relation in order to ensure that the unit which would be most important in a summary of the text is selected as the nucleus. However, this could equally well be seen as an argument for placing less emphasis on centrality and more emphasis on syntactic structure. In fact, Stede acknowledges that what is central to one's purposes can be very different from case to case, an observation that points to an important weakness in the notion of centrality. Incidentally, the notion of semantic prominence has been given up as a main criterion for headedness in syntax: syntactic theories routinely assume that main verbs may function as complements of auxiliaries and modals, although semantic prominence would seem to argue for the converse analysis.

In conclusion, it seems that many of the examples that have been suggested as counter-evidence against a tree-based discourse analysis that spans the entire discourse, can be accommodated within a syntax-centered discourse framework with a flexible compositional semantics and a rich set of mechanisms, e.g., for dealing with anaphoric discourse adverbials. In our view, the most intriguing outcome of this discussion is Stede's observation that the syntactic structure sometimes differs significantly from what is central to the writer's purposes. We could draw the conclusion that syntactic structure is less important than centrality, but the reverse

conclusion is just as possible: that it is the syntactic structure which is the more important of the two, which seems to be the near-universal conclusion in syntax.

7 Ambiguity and other remaining problems in the syntax-based view

A syntax-centered discourse annotation solves a number of problems – in particular, it allows the discourse to be represented as a tree with additional relations for coreference. But it also introduces some problems as well. Most importantly, whenever a nucleus has more than one adjunct, we can only compute the functor-argument structure if we are given a modifier scope. More generally, if we impose a highly principled linguistic framework on our annotation of discourse, including a tree-based model, it is obviously difficult to use these data to argue for the particular assumptions. This is however not a crucial objection, since exactly the same thing could be said about syntactic annotation.

8 The syntax-based discourse annotation in the CDT treebanks

The kind of syntax-centered discourse annotation we have described in the paper is currently being implemented in the Copenhagen Dependency Treebanks (CDT) to create a set of open-source parallel treebanks for five different languages, Danish, English, German, Italian, and Spanish (cf. Buch-Kromann et al., 2009; Buch-Kromann and Korzen, 2010). These treebanks resemble the Potsdam Commentary Corpus (Stede 2008) in that they are multi-level treebanks, i.e., the annotation includes syntactic structure, discourse structure, and coreference structure. The annotation is in its early stages, but more than 273 text excerpts with approximately 250 words in each excerpt have been annotated so far, using a detailed inventory of 50 discourse relations organized in a hierarchy so that different levels of granularity can be selected. The current inter-annotator agreement is approximately 50%, a number that we hope to improve.

9 Conclusions

The important question that we have sought to answer in this paper is whether discourse structure and syntactic structure are fundamentally different structures, or whether they are better viewed as instances of a single unified syntax-discourse tree structure at different levels of granularity in the segmentation. We have argued that if we think of discourse structure as an extended syntactic structure with an induced, but not explicitly expressed semantic predicate-argument structure that links the sentences in the entire discourse, the second, unified view is not only feasible, it also solves a number of syntax-discourse interface problems and provides a unified view of syntax and discourse that should make it easier to extend almost-linear parsing algorithms like the Malt parser (Nivre, 2006) to discourse parsing.

We have also argued that a syntax-centered view of discourse involves a significant departure from the original definition of nuclearity in Rhetorical Structure Theory, which is based on the notion of centrality to the writer's purposes, to a much more surface-oriented view of discourse structure. One possible concern is that discourse-based tasks like text summarization may become much harder in a syntax-

centered conception of discourse; on the other hand, discourse parsing might become easier because the resulting analyses are closer to the syntactic analyses.

References

Berfin Aktaş, Cem Bozşahin, Deniz Zeyrek, 2010. *Discourse Relation Configurations in Turkish and an Annotation Environment.* Proc. Linguistic Annotation Workshop (ACL-2010).

Matthias Buch-Kromann. 2009. *Discontinuous Grammar. A dependency-based model of human parsing and language learning.* VDM Verlag.

Matthias Buch-Kromann, Iørn Korzen & Henrik Høeg Müller. 2009. Uncovering the 'lost' structure of translations with parallel treebanks. *Copenhagen Studies in Language* 38: 199-224.

Matthias Buch-Kromann and Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. Proc. Linguistic Annotation Workshop (ACL-2010).

Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.

Lynn Carlson, Daniel Marcu, Mary Ellen Okurowski. 2001. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*. Proc. Discourse and Dialogue.

Noam Chomsky. 1981/1993. *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter.

Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Bonnie Webber, Aravind Joshi. 2002. *The discourse anaphoric properties of connectives.* Proc. DAARC 2002, pages 45-50.

William Croft. 1993. *What is a head*? In J. Rooryck and L. Zarin (eds.), *Phrase structure and the lexicon*, pages 35–76. Dordrecht: Kluwer Academic Publishers.

Mary Dalrymple, Ronald M. Kaplan, John Maxwell III, Annie Zaenen (eds.). 1994. *Formal issues in Lexical-Functional Grammar.* CSLI Lecture Notes, no. 47.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, Bonnie Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. *Proc. of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky,* pages 29-36.

David R. Dowty. 1992. Towards a minimalist theory of syntactic structure. In: Wietske Sijtsma and Arthur van Horck (eds.), *Discontinuous constituency*, Mouton de Gruyter.

Denys Duchier. 2001. Topological dependency trees: A constraint-based account of linear precedence. Proc. ACL-2001.

Katherine Forbes-Riley, Bonnie Webber, Aravind Joshi. 2006. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics*, 23(1).

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy* 14:39-100.

Richard Hudson. 1987. Zwicky on heads. *Journal of Linguistics* (23): 109-132.

Richard Hudson. 2010. *An introduction to Word Grammar.* Cambridge University Press.

Aravind Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In: Grzegorz Rozenberg and Arto Salomaa (eds.), *Handbook of Formal Languages. Beyond Words.* Springer-Verlag.

Alistair Knott. 2007. Review of 'Coherence in natural language: Data structures and applications', by Florian Wolf and Edward Gibson. *Computational Linguistics* 33:591–595.

Alex Lascarides and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure. In: Harry Bunt and Reinhard Muskens (ed.), *Computing Meaning.* Synthese language library, vol. 83. Springer Netherlands, pages 87-124.

Alan Lee, Rashmi Prasad, Aravind Joshi, Bonnie Webber. 2008. *Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank.* Proc. Constraints in Discourse III Workshop.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Bonnie Webber. 2006. *Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex Than in Syntax?* Proc. Treebanks and Linguistic Theories. Prague, Czech Republic. December 2006

William C. Mann and Sandra A. Thompson 1987. *Rhetorical Structure Theory. A Theory of Text Organization*. ISI: Information Sciences Institute, Los Angeles, CA, ISI/RS-87-190, 1-81.

Christopher D. Manning. 1995. Dissociating functor-argument structure from surface phrase structure: the relationship of HPSG Order Domains to LFG. Ms., Carnegie Mellon University.

Daniel Marcu. 2003. Discourse structures: trees or graphs http://www.isi.edu/~marcu/discourse/Discourse%20structures.htm

Igor Mel'čuk. 1988. *Dependency syntax.* State University of New York Press.

Lucie Mladová, Šarka Zikánová, Eva Hajičová. 2008. *From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank.* Proc. LREC-2008.

Joakim Nivre. 2006. *Inductive Dependency Parsing.* Springer.

Livia Polanyi. 1988. A Formal Model of Discourse Structure. Journal of Pragmatics 12: 601-639.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven phrase structure grammar.* Univ. of Chicago Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proc. 6th Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco.

Petr Sgall, Eva Hajičová, Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel Publishing Company.

Manfred Stede. 2008. RST revisited: Disentangling nuclearity. In: C. Fabricius-Hansen, W. Ramm (Eds.): '*Subordination' versus 'coordination' in sentence and text - A cross-linguistic perspective.* Studies in Language Companion Series. Amsterdam: John Benjamins.

Mark Steedman. 2000. The syntactic process. A Bradford Book, The MIT Press.

Bonnie Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28: 751-779.

Florian Wolf and Edward Gibson 2005. Representing Discourse Coherence: A Corpus-Based Study. Computational Linguistics 31(2), 249-287.

On the Dimensions of Discourse Salience

Christian Chiarcos
Universität Potsdam

Abstract

This paper describes results of two corpus studies of information packaging of discourse referents in German dedicated to the following questions:

- Do sentence-initial position, pronominalization and subject role assignment reflect a single underlying dimension of discourse salience or multiple dimensions ?
- If there are multiple dimensions of salience, is it possible to associate them with a forward-looking and a backward-looking perspective on discourse, as proposed, e.g., in the context of Centering (Grosz et al., 1995) ?

This paper presents empirical findings from TüBa-D/Z, a corpus of German newspaper articles, that provide evidence against a unidimensional model of discourse salience, and support the claim that (at least) two dimensions of salience are to be distinguished and that these dimensions are associated with different temporal orientations on discourse.

1 Salience and Information Packaging

In the last 30 years, the notion of "salience" has been employed in many accounts for the information packaging of discourse referents, especially with respect to the choice of referring expressions (e.g., realization as definite NP or as a pronoun), and with respect to the pragmatic function of grammatical roles and word order preferences: Personal pronouns are assumed to represent more salient referents than nominals (Sgall et al., 1986; Ariel, 1990; Grosz et al., 1995), the left periphery of sentences (and in particular, the sentence-initial position) are associated with a high degree of salience (Sgall et al., 1986; Sridhar, 1988; Rambow, 1993), and the grammatical subject is assumed to serve a similar function (Fillmore, 1977; Tomlin, 1995; Grosz et al., 1995).

Consider the German example sentence in (1). The expression *sie* 'they' is a subject pronoun in preverbal (*vorfeld*) position; according to the aforementioned theories, its referent is to be regarded as highly salient, whereas the nominals *auf einem Tandem* 'on a tandem' and *ins Stadion* 'to the stadion' are postverbal non-subjects and thus non-salient.

(1) Sie wollen auf einem Tandem ins Stadion radeln
they want.to on a tandem into.the stadion go.by.bike
'They want to go to the stadion by tandem.' (TüBa-D/Z, sentence 113)

Despite the apparent agreement on the relevance of salience to different information packaging phenomena, researchers disagree on determinants and the actual nature of salience. Already Sridhar (1988, p.38) noted that 'a number of different factors have been claimed to contribute to salience', that '[r]esearchers are (...) divided on the effects

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 31-44, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

of salience to sentences', and further that 'salience is obviously (...) characterized by a number of superficially dissimilar properties'.

Since then, three major views on the nature of salience have been established:

unidimensional: In traditional unidimensional models as advocated by Sgall et al. (1986, word order and referring expressions), Gundel et al. (1993) and Ariel (1990, referring expressions), and Tomlin (1995, word order and grammatical roles), salience is seen as **a single dimension of cognitive states**: Every referent is assigned a particular degree of salience and this degree of salience determines the packaging preferences for this referent.

multifactorial: The radical antithesis to the unidimensional view is to abandon the idea of a generalized notion of salience, and to focus on the study of **individual factors**. This has been the premise of the psycholinguistic research of Osgood and Bock (1977) and Sridhar (1988), and has been recently revived by Kaiser and Trueswell (2004, to appear 2011) and Brown-Schmidt et al. (2005).

multidimensional: Multidimensional models of salience postulate the existence of multiple dimensions of salience as independent generalizations over a certain range of different factors. Typically, two dimensions are distinguished (Givón, 1983; Pattabhiraman and Cercone, 1990; Clamons et al., 1993; Mulkern, 2007): One dimension that is primarily defined with respect to the preceding discourse and/or the common ground, and that is thus primarily **backward-looking**. The other dimension is more concerned with the intentions and goals of the speaker and takes into consideration how these are manifested in subsequent discourse, and that is thus (at least partially) **forward-looking**.[1]

Psycholinguistic experiments and corpus studies on personal pronouns and demonstrative pronouns conducted by Kaiser and Trueswell (2004, to appear 2011), Brown-Schmidt et al. (2005) and Ellert and Hopp (2010) indicate that personal pronouns and demonstrative pronouns deviate in their antecedent selection preferences (in Finnish, Estonian, Dutch, English and German). In other words, certain salience factors contribute *independently* to the choice and interpretation of referring expressions in these languages. This observation can be seen as direct counterevidence for a unidimensional model that would postulate that demonstrative pronouns and personal pronouns reflect cognitive states organized in one uniform dimension of salience.

However, it is not necessary to conclude that multiple *cognitive* dimensions are involved: A functionalist with a unidimensional salience model might argue that the specific preference of personal pronouns to take subject antecedents can be attributed

[1] A representative multidimensional model of salience is the proposal of Clamons et al. (1993) and Mulkern (2007). They postulate that every referent is characterized by the degree of salience arising from the preceding context ('givenness', 'inherent salience') on the one hand, and on the other hand by the degree of salience imposed on this referent by the speaker ('importance', 'emphasis', 'imposed salience') in order to increase its accessibility in subsequent discourse. Similar approaches (with different terminologies) have been described by Givón (1983, 2001); Pattabhiraman and Cercone (1990), and are also suggested by Levelt (1989) and Chafe (1994).

to grammaticalization tendencies (comparable to those that lead to the development of syntactically bound (e.g., relative) pronouns out of anaphoric demonstrative pronouns in German, see Diessel, 1999, p.120ff), and further that these grammaticalization tendencies are actually based on conventional patterns of salience.[2] From a functional point of view, differences in subject-sensitivity are thus no counterevidence to a unidimensional salience model, because it is not only salience that affects packaging preferences but also grammatical conventions (that rely on the same conception of salience). This line of argumentation may be applied to every attempt to prove the insufficiency of unidimensional models of salience by showing that different types of referring expressions (or grammatical roles etc.) differ in their sensitivity to specific salience factors, especially those that address the linguistic realization of the antecedent.

In order to distangle grammaticalization tendencies from salience, it is thus necessary to evaluate predictions of unidimensional models of salience independently from the study of individual salience factors. This is the aim of the corpus study described in Sec. 3: Independently from the salience factors involved, a unidimensional model of salience predicts correlations between preverbal word order, pronominalization and subject role assignment. If the expected correlation between these phenomena cannot be confirmed, we have to conclude that at least two different dimensions or factors must be involved in the information packaging of these phenomena.

As an alternative to unidimensional models, Kaiser and Trueswell (2004, to appear 2011) suggest a multifactorial approach. A factor-based model, however, misses an important generalization, i.e., a theoretically motivated explanation for the underlying processes involved in information packaging, cf. the early critical remarks by Tomlin (1995). From a theoretical point of view – but also from the perspective of natural language processing (NLP) applications that try to interpret and reproduce information packaging preferences –, it is thus desirable to abstract from individual factors. Multidimensional models of salience provide such an abstraction in that they propose a dichotomy of logically independent dimensions of salience that interact in the process of information packaging. The second corpus study (Sec. 4) addresses the question whether these dimensions of salience correlate with forward-looking and backward-looking functions of referring expressions in discourse.

2 Corpus and Feature Extraction

The corpus studies described below are conducted on non-coordinated main clauses from the TüBa-D/Z corpus (v.5), a corpus of 2,213 German newspaper articles annotated for morphology, syntax and coreference (Telljohann et al., 2009; Naumann,

[2] In a unidimensional model of salience, grammatical roles and pronominalization are indirectly associated through the conception of salience: The subject represents the most salient referent of the current clause, and if we assume that this referent is most likely to remain the most salient referent of the following clause, it is to be expected to be realized as a personal pronoun then. By grammaticalization, the indirect association between subject role of the antecedent and the choice of a personal pronoun (originally mediated by the concept of salience) develops into a direct association, i.e., a grammatical convention that the original personal pronoun takes a subject antecedent.

2007). TüBa-D/Z is particularly well-suited for this study, as it combines anaphoric annotations with explicit annotations of topological fields of German sentences. With respect to word order, we can thus make use of the theoretically well-founded concept of *vorfeld* constituents.

From the corpus, all non-coordinated, non-embedded main clauses (40,713 clauses) were extracted, and all their nominal and pronominal arguments and adjuncts[3] were considered as (potential) referring expressions (79,222 in total).

The following classes of referring expressions (`ref`) were distinguished:

- 6 types of pronominal expressions
  - `perspron` personal pronoun, `pronadv` pronominal adverb, `dempron` demonstrative pronoun, `reflpron` reflexive pronoun, `pron` other pronouns (e.g., pronominal quantifiers)
- 7 types of nominal expressions
  - `name` proper name, `defNP` definite NP, `indefNP` indefinite NP[4], `possNP` possessive NP, `demNP` demonstrative NP, `NP` other NPs (e.g., NPs with semidemonstrative determiner *solch* 'such', or interrogative determiner *welch* 'which')
- coordinations
  - `coord,pron` pronominal coordination (all conjuncts are pronominal), `coord,NP` nominal coordination (at least one conjunct is nominal)

Depending of the parent nodes of the expressions under consideration in the syntax annotation, four different word order possibilities (`wo`) were distinguished.

- `vf` *vorfeld* positioning, preverbal (node label `VF`),
- `mf_initial` *mittelfeld* initial, immediately after the finite verb (NP/PP that is the *left-most* child of an `MF` node),
- `mf_noninitial` in the *mittelfeld*, but preceded by another expression,
- `nf` *nachfeld*, a right-peripheral field, following displaced verbal particles and infinite verbs (node label `NF`).

Three classes of grammatical roles (`gr`) were distinguished:

- `sbj`, grammatical subject (edge labels `on`, `onk`)
- `obj`, non-prepositional object (edge labels `oa`, `oak`, `od`, `odk`, `og`, `ogk`)
- `other`, remaining complements (incl. prepositional objects, other PPs, predications, etc.)

The values of `ref`, `wo` and `gr` represent the packaging phenomena distinguished in the first corpus study. For the second corpus study, two additional features, `given` and `important`, were derived from the coreference annotation:

- `given`, if linked to another expression in the preceding discourse by `coreferential`, `anaphoric`, `bound`, `cataphoric` or `instance` relations.
- `important`, if linked to another expression in the subsequent discourse by `coreferential`, `anaphoric`, `bound`, `cataphoric` or `instance` relations.

[3] `NX` and `PX` nodes directly attached to `VF`, `MF`, or `NF`

[4] As defined here, indefinite also includes determinerless and quantified NPs.

Feature extraction was performed using an extension of Gerlof Bouma's Prolog interface to TüBa-D/Z (Bouma, 2010).[5]

3 Predictions of the Unidimensional Model

The first corpus study investigates the predictions of the unidimensionality hypothesis with respect to expected correlations between *vorfeld* positioning (`vf`), subject role assignment (`sbj`) and pronominalisation (`perspron`) of referring expressions in German main clauses.

If *vorfeld* positioning, subject role assignment and pronominalisation all serve as indicators of a particularly high degree of a single dimension of salience, then `sbj` entails a particularly highly salient referent, this referent is thus to be represented as `perspron` (with respect to referring expressions) and `vf` (with respect to word order). In terms of conditioned probabilities, a unidimensional model of salience entails the following inequations:

$$P(\texttt{perspron}|\texttt{sbj}) \;>\; P(\texttt{perspron}) \tag{1}$$

$$P(\texttt{vf}|\texttt{sbj}) \;>\; P(\texttt{vf}) \tag{2}$$

or, more generally, for any two grammatical devices $X^{sal\uparrow}$ and $Y^{sal\uparrow}$ that are associated with particularly high degrees of salience:

$$P(X^{sal\uparrow}|Y^{sal\uparrow}) \;>\; P(X^{sal\uparrow}) \tag{3}$$

This formulation of the unidimensional model relies on the following additional assumptions:

(1) Pronominalization, subject role assignment and placement in the *vorfeld* are determined by grammatical (syntactic/semantic) and functional determinants.

(2) There is no semantic or syntactic constraint that disencourages the cooccurence of pronominalization, subject role and *vorfeld* positioning.

(3) Salience is the primary functional determinant of pronominalization, subject role and *vorfeld* positioning.

(4) Pronominalization, subject role and *vorfeld* positioning indicate high degrees of salience.

Assumption (1) states that the impact of language-independent factors on information packaging, e.g., biological factors, is marginal as compared to the impact of functional and grammatical factors. This is the fundamental assumption underlying the existing

[5]The code developed for this purpose is available under `http://www.ling.uni-potsdam.de/~chiarcos/` under "Resources".

literature on salience in discourse. Assumption (2) expresses the fact that a sentence as in ex. (1) is both syntactically well-formed and semantically felicitous. Assumption (3) is an assumption of any salience-based account of information packaging; (4) represents generally accepted claims on the impact of salience on information packaging.

Under these assumptions, a unidimensional model of salience predicts correlation between pronominalization, subject role assignent and *vorfeld* positioning, as the assumptions (1) to (4) entail that causal relationships between these packaging phenomena that are not mediated by salience are marginal if not inexistent.

As stated in assumptions (1) and (3), information packaging is, however, not exclusively determined by salience, but other factors may also play a role, although to a lower degree than salience: Subject role assignment is influenced, for example, by animacy. Also, word order preferences and pronominalization are affected by other factors besides salience. With one underlying dimension of salience, however, the effects of such circumstantial factors can be minimized if only those referents are considered that are marked as being salient with respect to *two* dimensions of information packaging, e.g., a referent that is both subject and in *vorfeld*. The expected minimization of such circumstantial factors can be captured in the following inequations:

$$P(\texttt{perspron}|\texttt{sbj},\texttt{vf}) \;\geq\; P(\texttt{perspron}|\texttt{sbj}) \tag{4}$$

$$P(\texttt{perspron}|\texttt{sbj},\texttt{vf}) \;\geq\; P(\texttt{perspron}|\texttt{vf}) \tag{5}$$

or, more generally

$$P(X^{sal\uparrow}|Y^{sal\uparrow}, Z^{sal\uparrow}) \;\geq\; P(X^{sal\uparrow}|Y^{sal\uparrow}) \tag{6}$$

Inequations (3) and (6) represent predictions that immediately follow from a unidimensional model of salience under the assumptions given above. If they cannot be confirmed in the data, then either one of the packaging phenomena under consideration is not actually a salience-indicating grammatical device (despite the support from the literature), or a unidimensional model of salience is inappropriate for the formalization of information packaging for the choice of referring expressions, the assignment of grammatical roles and word order preferences at the same time.

Table 1 summarizes the results obtained for inequation (3). For all grammatical devices, we can observe an increase of relative frequency under the condition of another salience-marking grammatical device as entailed by the unidimensionality hypothesis. This indicates that there is indeed a functional overlap between `perspron`, `sbj` and `vf`.

However, the marginal increase of `perspron` probability under the condition `vf` (and vice versa) may rise suspicions that the functional overlap between these three

realization $X^{sal\uparrow}$	condition $Y^{sal\uparrow}$	(conditioned) probability $P(X^{sal\uparrow}\|Y^{sal\uparrow})$	probability increase (vs. unconditioned)
perspron	(none)	10.80% (8,557/79,222)	
	vf	11.43%	+0.63%
	sbj	20.06%	+9.26%
sbj	(none)	42.50% (33,667/79,222)	
	perspron	78.94%	+36.44%
	vf	63.91%	+21.41%
vf	(none)	33.16% (16,789/79,222)	
	perspron	35.08%	+1.92%
	sbj	49.87%	+16.71%

Table 1: $P(X^{sal\uparrow}|Y^{sal\uparrow}) > P(X^{sal\uparrow})$ in TüBa-D/Z ?

realization		χ^2	ϕ
±perspron	±vf	p < .0001	.014
±perspron	±sbj	p < .0001	.257
±sbj	±vf	p < .0001	.305

Table 2: Significance (χ^2) and Pearson correlation coefficient (ϕ) of perspron, sbj, and vf

phenomena is actually a functional overlap between sbj and perspron on the one hand and between sbj and vf on the other hand, while vf and perspron are only loosely related. Nevertheless, also the latter correlation is highly significant and positive for all pairs of packaging phenomena considered, as shown in Table 2.[6]

While the corpus data does not contradict the predictions of (3), inequation (6) could not be confirmed: Against the expected increase of probability under the condition of two salience-marking grammatical devices as compared to a single salience-marking grammatical device, Table 3 shows a **decrease** of probability for subject pronouns in *vorfeld* (unlike pronouns under the condition of being subject, and *vorfeld* under the condition of being subject):

$$P(\texttt{perspron}|\texttt{vf},\texttt{sbj}) \;<\; P(\texttt{perspron}|\texttt{sbj}) \tag{7}$$

$$P(\texttt{vf}|\texttt{perspron},\texttt{sbj}) \;<\; P(\texttt{vf}|\texttt{sbj}) \tag{8}$$

Apparently, there is a bias against subject pronouns in *vorfeld* (albeit there is no evidence for a bias against pronouns **or** subjects in *vorfeld*). This is a clear violation of predictions of the unidimensionality hypothesis. As we excluded circumstantial factors and grammatical well-formedness conditions as potential causes for divergency, we have to conclude that there are (at least) two functional dimensions underlying pronominalization, subject role assignment and *vorfeld* positioning, and further, that

[6] In Table 2 and in the remainder of this paper, ±X means that X applies (e.g., +sbj) or that X does not apply (e.g., −sbj that matches obj and other).

realization	conditions		probability	probability increase vs.	
$X^{sal\uparrow}$	$Y^{sal\uparrow}$	$Z^{sal\uparrow}$	$P(X^{sal\uparrow}\|Y^{sal\uparrow}, Z^{sal\uparrow})$	$P(X^{sal\uparrow}\|Y^{sal\uparrow})$	$P(X^{sal\uparrow}\|Z^{sal\uparrow})$
perspron	vf	sbj	15.51% (2,604/16,789)	+4.08%	**−4.55%**
vf	perspron	sbj	38.55% (2,604/6,755)	+3.47%	**−11.32%**
sbj	vf	perspron	86.74% (2,604/3,002)	+22.84%	+7.80%

Table 3: $P(X^{sal\uparrow}|Y^{sal\uparrow}, Z^{sal\uparrow}) \geq P(X^{sal\uparrow}|Y^{sal\uparrow})$ in TüBa-D/Z ?

subject role assignment is associated with both dimensions.

As for contextual features involved with these dimensions, the structure of the preceding discourse is generally assumed to play an important role: Previous mention and the linguistic realization of the antecedent are commonly considered to be a major determinant of pronominalization (Sgall et al., 1986; Ariel, 1990; Grosz et al., 1995), it is assumed to be associated with subject role assignment (Prince, 1992; Lambrecht, 1994, also cf. preference for continue transitions in Grosz et al., 1995), and traditionally with *vorfeld* positioning as well (the original working hypotheses of Speyer, 2007, and Dipper and Zinsmeister, 2009). A number of recent corpus studies, however, could not confirm that the preceding discourse determines *vorfeld* positioning, and a number of alternative factors have been suggested in consequence:

- Evidence against the primarily anaphoric nature of the *vorfeld* can be drawn from a number of recent corpus studies that actually attempted to *prove* the relevance of the preceding context to *vorfeld* positioning: For a collection of German prose text from various genres, Speyer (2007) reported that 51% of *vorfeld* constituents could be neither semantically nor anaphorically linked to the preceding discourse. On a corpus of parliamentary debates, Dipper and Zinsmeister (2009) found that 55% of *vorfeld* constituents stand in no obvious relationship to the preceding discourse, whereas only 23% are anaphorically linked. In their study of object arguments in the *vorfeld* of main clauses in the NEGRA corpus, Weber and Müller (2004) found no indication that anaphoric (given/definite/pronominal) objects tend to precede non-anaphoric (new/indefinite/nominal) subjects. In fact, indefinite objects preceded definite subjects in OVS sentences more often than vice versa.

- A smaller number of corpus studies and theoretical papers have dealt with alternative factors contributing to *vorfeld* positioning: For example, Filippova and Strube (2007) found that *vorfeld* constituents tend to refer to the global discourse topic (i.e., names mentioned in the headlines in their collection of biographic articles). Speyer (2007) proposed that the *vorfeld* is the preferred locus of contrastive ex-

pressions and frame-setting topics, whereas purely anaphoric expressions are positioned there only if the *vorfeld* would have been left unoccupied otherwise (cf. Frey (2004a) for a similar model).

Previously proposed non-anaphoric factors of *vorfeld* positioning often involve certain intentions on the side of the speaker, i.e., to express contrastivity, importance or to make sure that subsequent of information are interpreted in the context of a particular situational environment. In the words of Lötscher (1984), these functions may be seen as specific aspects of the 'highlighting' function of the *vorfeld*.[7]

Of course, it is problematic to quantify 'highlighting' without having direct access to the mental discourse model of the speaker at the moment of uttering. But at least one aspect of 'highlighting' can be extrapolated from the text itself – the speaker's intention to prepare the hearer for the forthcoming discourse: By placing the referent in preverbal position (or by choosing an otherwise prominent realization), the speaker performs a 'foregrounding' operation whose effects on the subsequent discourse can be observed, in particular, the increased anaphoric accessibility of the referent. The speaker's foregrounding intentions can thus be inferred from the distribution and the realization of the referent in the forthcoming discourse. As far as foregrounding is concerned, 'highlighting' can thus be approximated by forward-looking factors.[8]

The second corpus study evaluates the hypothesis that the dimensions of salience involved in *vorfeld* positioning, subject role assignment and pronominalisation can be aligned with such a backward-looking/forward-looking dichotomy.

4 Temporal Dimensions of Salience

The dichotomy between forward-looking and backward-looking salience is adopted in most multidimensional models of salience (Givón, 1983, 2001; Clamons et al., 1993; Mulkern, 2007, see also Grosz et al., 1995), and the corpus study described in this section investigates the relevance of this distinction for the packaging phenomena under investigation here.

As maximally theory-independent metrics of salience, backward-looking salience is reduced here to the existence of a previous reference to the same referent (`+given`), forward-looking salience is approximated by the existence of a subsequent mention of the same referent (`+important`).

A series of χ^2 square tests where the features `±perspron`, `±sbj` and `±vf` were tested against `±given` and `±important` reveals a significant interaction and a pos-

[7] Alternative terms include 'newsworthiness' (Mithun, 1992), 'imposed salience' (Clamons et al., 1993; Mulkern, 2007), or 'importance' (Givón, 1988), all discussed with respect to word order inverse and fronting in multiple languages.

[8] While this does not mean that forward-looking salience can be **equated** with the speaker's intention to highlight certain referents, forward-looking factors allow to reconstruct a certain fraction of the speaker's intentions at the moment of utterance, but only those that deal with his intention to prepare the hearer for the development of the subsequent discourse in order to guide him to a specific insight. That conversation involves such antipacitory elements was already emphasized by Grosz et al. (1995) who mention that speakers 'should plan ahead to minimize the number of shifts'. Of course, other intentions of the speaker, e.g., emotions or the intention to trigger certain implicatures (Ariel, 1990; Gundel et al., 1993) cannot be reconstructed with this heuristic. Every approximation of the speaker's original intentions by means of forward-looking factors is thus incomplete, but nevertheless feasible with respect to the foregrounding function of grammatical devices.

realization	±given χ^2	±given ϕ	±important χ^2	±important ϕ
±perspron	$p < .0001$	.342	$p < .0001$	.174
±sbj	$p < .0001$	.288	$p < .0001$	.279
±vf	$p < .0001$	.065	$p < .0001$	.073

Table 4: Significance (χ^2) and Pearson correlation coefficient (ϕ) of ±given/ ±important and packaging phenomena

itive correlation between the packaging phenomena considered and both dimensions of salience (Table 4).

In order to assess how ±given and ±important interact during the derivation of packaging preferences, C4.5 decision trees were trained on the feature sets extracted from TüBa-D/Z (using the J48 implementation of WEKA, Witten and Frank, 2005): The C4.5 algorithm maximizes the correctness of classification, and with ±given and ±important as input features and different packaging phenomena as target classification, the decision tree built up by algorithm allows to extrapolate the influence of previous mention and of subsequent mention on the choice of referring expressions, the assignment of grammatical roles and word order preferences.

(a) referring expressions (ref)

```
correctness: 34.6%
(baseline: defNP, 33.6%)

+given
|   +important: perspron
|   -important: defNP
-given: defNP
```

(b) grammatical roles (gr)

```
correctness: 53.1%
(baseline: sbj, 42.5%)

+given: sbj
-given
|   +important: sbj
|   -important: other
```

(c) word order (wo)

```
correctness: 38.7%
(baseline: mf_initial, 33.6%)

+given: mf_initial
-given
|   +important: vf
|   -important: mf_noninitial
```

Figure 1: C4.5 decision trees to predict packaging preferences from ±given and ±important.

For every dimension of information considered here, ref, gr and wo, a decision tree was trained to predict the actual grammatical device (with the fine-grained subclasses as described in Sec. 2) based on the features ±given and ±important. The resulting trees are shown in Fig. 1. Compared with the baseline (most frequent class), all classifiers yield an increase in correctness.[9] A more interesting evaluation method is a comparison of the classifier strategies with claims in linguistic literature:

[9]Note that these classifiers only serve as an indicator of the way that ±given and ±important influence information packaging. The overall classification results are poor, but mostly because the number of packaging phenomena distinguished for the different levels of information packaging is far greater than the number of possible combination of input values. However, with more fine-grained measurements of backward-looking and forward-looking salience, as studied, for example, by Chiarcos (2009), more detailed information packaging predictions may be possible. At this point, such improvements are left as a topic for subsequent research.

We can observe a remarkable degree of compatibility of the classifiers with theoretical models.

- As expected from the literature, the `ref` classifier predicts a close association between `given` and `perspron`. That `important` has an impact on pronominalization may reflect a preference to maintain an established (pronominal) topic over several utterances, cf. Lambrecht (1994, p.199ff.), Grosz et al. (1995).
- The `gr` classifier combines two conflicting views on functional determinants of subjecthood found in two different branches in the literature: Traditionally, the subject is associated with high degrees of backward-looking salience (e.g., Prince, 1992), but in typological literature, it is assumed that subjects serve an attention-guiding, foregrounding function (Tomlin, 1995; Pustet, 1997).[10] The classifier combines both views by stating that the subject is important *or* given.
- The `wo` classifier closely resembles modern approaches on *vorfeld* positioning in German: Frey (2004a,b) postulated that the unmarked position of the (given-ness-) topic is the immediate post-verbal position (the preferred locus of *given* referents according to the `wo` classifier), whereas placement of the topic in the *vorfeld* requires the presence of an another pragmatic force, e.g., 'kontrast' as defined by Vallduví and Vilkuna (1998). Above, a functional resemblance between foregrounding and contrast was suggested (both represent different aspects of the 'highlighting' force of the *vorfeld*), so that the preference to place subsequently mentioned referents in `vf` can be compared to the effect of kontrast in Frey's model.

  The predicted effects of the feature bundles `+given` and `−given/+important` on `wo` can also be compared to Lambrecht's information-structural characterization of the left periphery: Lambrecht (1994, p.199ff) assumes that the sentence-initial position serves the function of *topic announcement*, i.e., that a referent that has not been established as a topic before (`−given`) is marked as being the topic of the subsequent discourse segment (`+important`). As opposed to this, the function of *topic maintenance* of already established (`+given`) topics is not associated with the left periphery, but with proximity to the finite verb, i.e., `mf_initial` in German main clauses.

As mentioned above, decision trees can be rephrased as rules for information packaging preferences and thus compared with regularities reported in the linguistic literature. Table 5 summarizes these rules and reveals another remarkable coincidence: `±given` and `±important` predict exactly the distribution of grammatical devices observed in the first corpus study:

[10] One of the few models that combines both views can be found in Centering (Grosz et al., 1995) where subjects are ascribed both a forward-looking function (the subject is the preferred center, i.e., highest-ranking forward-looking center), and a backward-looking function (preference for identity of preferred center and backward-looking center, preference of continuity of the backward-looking center).

±given	±important	prediction		
+	+	perspron	sbj	mf_initial
+	−	defNP	sbj	mf_initial
−	+	defNP	sbj	vf
−	−	defNP	other	mf_noninitial

Table 5: Information packaging preferences predicted from ±given and ±important

(a) an association between pronominalization and subject role (+given, +important),

(b) an association between *vorfeld* positioning and subject role (−given, +important), and

(c) a dispreference for subject pronouns (+given) to coincide with *vorfeld* (−given).

5 Results

Taken together, both corpus studies provide evidence against a unidimensional model of discourse salience, and, more specifically, they support the claim that (at least) two dimensions of discourse salience are to be distinguished, and further that these dimensions are associated with different temporal orientations on discourse:

- It is necessary to distinguish at least two dimensions of salience in order to account for *vorfeld* positioning, pronominalization and subject role assignment in a salience-based model.
- Previous mention (backward-looking salience, givenness) and subsequent mention (forward-looking salience, importance) have a highly significant influence on these packaging phenomena.
- The interaction between both factors leads to the observed distribution of these packaging phenomena.

Acknowledgements

The research described in this paper was was funded by the German Research Foundation (DFG) in the context of the Collaborative Research Center (SFB) 632, Project D1 "Linguistic Database" at the Universität Potsdam, Germany. I would like to thank three anonymous reviewers, Stefanie Dipper, Heike Zinsmeister, Julia Ritz and Robin Hörnig for comments and feedback, and Gerlof Bouma for his support during the development of the feature extraction scripts.

References

Mira Ariel. *Accessing Noun-Phrase Antecedents*. Routledge, London, New York, 1990.

Gerlof Bouma. Syntactic tree queries in Prolog. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV), held in conjunction with ACL 2010*, pages 212–216, Uppsala, Sweden, July 2010.

Sarah Brown-Schmidt, Donna K. Byron, and Michael K. Tanenhaus. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, 53:292–313, 2005.

Wallace Chafe. *Discourse, Consiousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago and London, 1994.

Christian Chiarcos. *Mental Salience and Grammatical Form. Toward a Model of Salience for Natural Language Generation*. PhD thesis, Universität Potsdam, Germany, Nov 2009.

C. Robin Clamons, Ann E. Mulkern, and Gerald Sanders. Salience signaling in Oromo. *Journal of Pragmatics*, 19:519–536, 1993.

Holger Diessel. *Demonstratives. Form, Function, and Grammaticalization*. John Benjamins, Amsterdam, Philadelphia, 1999.

Stefanie Dipper and Heike Zinsmeister. The role of the German vorfeld for local coherence: A pilot study. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 69–79, Narr, Tübingen, Sep 2009.

Miriam Ellert and Holger Hopp. Disentangling topicality from order of mention in the resolution of the German subject pronouns *er* and *der*: Off-line and on-line data. In *Proceedings of the Biennal Conference of the German Society of Cognitive Science (KogWis 2010)*, Potsdam, Germany, Oct 2010.

Katja Filippova and Michael Strube. The German vorfeld and local coherence. *Journal of Logic, Language and Information*, 16(4):465–485, 2007.

Charles J. Fillmore. Topics in lexical semantics. In Roger W. Cole, editor, *Current Issues in Linguistic Theory*, pages 76–138. Indiana University Press, Bloomington, 1977.

Werner Frey. The grammar-pragmatics interface and the German prefield. *Sprache und Pragmatik*, 52:1–39, 2004a.

Werner Frey. A medial topic position for German. *Linguistische Berichte*, 198:153–190, 2004b.

Talmy Givón. Introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, pages 5–41. John Benjamins, Amsterdam and Philadelphia, 1983.

Talmy Givón. The pragmatics of word order: Predictability, importance and attention. In Michael Hammond, Edith A. Moravcsik, and Jessica Wirth, editors, *Studies in Syntactic Typology*, pages 243 – 284. John Benjamins, Amsterdam and Philadelphia, 1988.

Talmy Givón. *Syntax*. John Benjamins, Amsterdam and Philadelphia, 2001.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Jeanette K. Gundel, Nancy A. Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):247–307, 1993.

Elsi Kaiser and John Trueswell. The referential properties of Dutch pronouns and demonstratives: Is salience enough? In Matthias Weisgerber and Cecile Meier, editors, *Proceedings of Sinn und Bedeutung 8*, University of Konstanz linguistics working papers, pages 137–150, Konstanz, 2004.

Elsi Kaiser and John Trueswell. Investigating the interpretation of pronouns and demonstratives in Finnish: Going beyond salience. In Edward Gibson and Neal J. Pearlmutter, editors, *The Processing and Acquisition of Reference*. MIT Press, Cambridge, Mass, to appear 2011.

Knud Lambrecht. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, 1994.

Willem J.M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, 1989.

Andreas Lötscher. Satzgliedstellung und funktionale Satzperspektive. In Gerhard Stickel, editor, *Pragmatik in der Grammatik*, Jahrbuch 1983 des Instituts für deutsche Sprache, pages 118–151. Schwann, Düsseldorf, 1984.

Marianne Mithun. Is basic word order universal? In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 15–62. John Benjamins, Amsterdam and Philadelphia, 1992.

Ann E. Mulkern. Knowing who's important: Relative discourse salience and Irish pronominal forms. In Nancy A. Hedberg and Ron Zacharski, editors, *The Grammar-Pragmatics Interface: Essays in honor of Jeanette K. Gundel*, pages 113–142. John Benjamins, Amsterdam and Philadelphia, 2007.

Karin Naumann. Manual for the Annotation of in-document Referential Relations. Technical report, Universität Tübingen, Seminar für Sprachwissenschaft, 2007. version of May 2007.

Charles E. Osgood and J. Kathryn Bock. Salience and sentencing: Some production principles. In Sheldon Rosenberg, editor, *Sentence Production: Developments in Research and Theory*, pages 89–140. Erlbaum, Hilldale, N.J., 1977.

T(hiyagarajasarma) Pattabhiraman and Nick Cercone. Selection: Salience, relevance and the coupling between domain-level tasks and text planning. In *Proceedings of the 5th International Workshop on Natural Language Generation (IWNLG 1990)*, pages 79–86, Pittsburgh, Apr 1990.

Ellen F. Prince. The ZPG letter: Subjects, definiteness, and information-status. In Sandra A. Thompson and William C. Mann, editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*, pages 295–325. John Benjamins, Amsterdam and Philadelphia, 1992.

Regina Pustet. *Diskursprominenz und Rollensemantik – Eine funktionale Typologie von Partizipantensystemen*. Lincom Europa, München, 1997.

Owen Rambow. Pragmatic aspects of scrambling and topicalization in German. In *Workshop on Centering Theory in Naturally-Occurring Discourse*. Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, 1993.

Petr Sgall, Eva Hajićová, and Jarmila Panevova. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Augustin Speyer. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 26(1):83–116, 2007.

Shikaripur N. Sridhar. *Cognition and Sentence Production. A Cross-Linguistic Study*. Springer, New York and Berlin, 1988.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, 2009. version of November 2009.

Russel S. Tomlin. Focal attention, voice, and word order. An experimental, cross-linguistic study. In Mickey Noonan and Pamela Downing, editors, *Word Order in Discourse*, pages 517–554. John Benjamins, Amsterdam and Philadelphia, 1995.

Enric Vallduví and Maria Vilkuna. On rheme and kontrast. In Peter Culicover and Louise McNally, editors, *The Limits of Syntax*, pages 79–108. Academic Press, New York, 1998.

Andrea Weber and Karin Müller. Word order variation in German main clauses: A corpus analysis. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora. Held in Conjunction with the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 71–78, Geneva, August 2004.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.

Annotating Information Structure: The Case of *Topic*

Philippa Cook & Felix Bildhauer
Freie Universität Berlin

Abstract
This paper deals with the annotation of Sentence Topics/Aboutness Topics in naturally occurring data. We report on a corpus study in which relatively poor inter-rater agreement was attained for the annotation of topics, although both coders were adhering to the same annotation instructions. Tokens that were particularly difficult to assess are identified, systematized, and discussed in some detail. In sum, the cases that are most likely to lead to non-matching annotations are those that either require a decision between "thetic" or "topic-comment", or involve an overlap between Focus and Topic. The findings raise a number of issues that may contribute to the discussion in theoretical linguistics, and they also may alert other researchers planning a similar enterprise to some pitfalls they may encounter.

1 Introduction
Research on information structure may serve a twofold purpose: first, information structure constitutes an intriguing area of investigation in its own right, where numerous concepts and their interrelations are still in need of further refinement. Second, insights in this field may lead to promising (re-)analyses of linguistic phenomena on the basis of information structure, that is, using information structural constraints in describing phenomena previously accounted for in terms of syntax (e.g. De Kuthy, 2002; Cook and Payne, 2006; Ambridge and Goldberg, 2008; Cook and Ørsnes, 2010). In this respect, corpora annotated for information structure are particularly valuable, as they put one in a position to test linguistic analyses that are based on notions such as "topic", "focus" and "givenness". However, not only are these notions used in different ways across different currents of research, but they also cause considerable problems when applied to naturally occurring data by researchers who otherwise agree largely on the definitions of these concepts and who even adhere to the same set of annotation guidelines.

In the present paper, we will take a closer look at the annotation of Sentence Topics/Aboutness Topics in naturally occurring data. The data we will discuss were extracted from the DeReKo[1] corpus and coded for information structure as part of a study on preferential topic realization in German newspaper texts, that is, a corpus study not initially related to the present work. Section 2 outlines the criteria used by the annotators for identifying Aboutness Topics and relates them to alternative approaches to topic-hood. Section 3 reports the relevant details of the corpus study, including measurements of inter-rater agreement for the annotation of Aboutness Topics. In Section 4, we identify the type of data that turns out to be particularly difficult to assess

[1] http://www.ids-mannheim.de/kl/projekte/korpora/

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena", pages 45-56, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

and seek to establish exactly which features are involved in these cases and why they give rise to diverging annotations.

2 "Topic" in theory and in annotation guidelines

The notion of Topic we are dealing with in the study being reported on here is that of Aboutness Topic. Since there are considerable differences in the way in which the notion of Topic has been used, and since the actual operationalizability of this notion is the crux of the current contribution, we will lay out here some of the basic assumptions taken by researchers working with the notion of Aboutness Topic. Krifka (2007), in his concise overview of the basic notions of information structure, points out that the use of the terms Topic and Comment reflect what von der Gabelentz (1869) called 'psychological subject' and 'psychological predicate' respectively, that is "the entity that a speaker identifies about which then information, the comment, is given " (Krifka, 2007, 40). This approach to Topic was further elaborated by Reinhart (1981), who adopts Stalnaker's (1978) notion of "context set" (a set of propositions which interlocutors accept to be true; that is, a "Common Ground"). In addition, Reinhart assumes that the Common Ground is structured in a such a way that information is stored in terms of a pairing of an entity and a proposition (or set of propositions) about that entity. New information is added to the Common Ground in the form of structured propositions, where the Sentence Topic designates an entity and the remainder of the sentence contributes the information to be associated with that entity (just like information in a file-card system is stored on a certain file card bearing a heading).[2] Building on Reinhart's approach, Krifka (2007) formulates the following definition:

(1) The topic constituent identifies the entity or set of entities under which the information expressed in the comment constituent should be stored in the CG content.

The notions of Topic/Comment have sometimes been mixed up with the notions of Background/Focus such that, for instance, Focus is believed to be the complement of Topic. The reason for this mixing of dimensions is presumably due to the fact that Topics are in practice prototypically given whereas, in contrast, foci are canonically new. Thus there seems to be a simple dichotomy in which newness and givenness align independently with Focus and Topic respectively. Such a merging of the dimensions is, however, problematic because there are cases which deviate from the canonical alignment in that (i) there are topics which contain a Focus, viz. e. g. (2) below. Such examples involving so-called contrastive topics can, but do not have to, involve an aboutness topic. Rather, the unifying feature of so-called contrastive topics is their function in discourse, where they are assumed to indicate a discourse strategy (Roberts, 1996; Büring, 2003; Krifka, 2007). The next problematic case is (ii) that there do appear to be new (i. e. non-given) topics as in (3) where an entity is introduced as new into the discourse

[2]The file-card metaphor has since been used by a number of authors. For a critical evaluation, see e. g. Hendriks and Dekker (1996).

(*a good friend of mine*) but still serves as aboutness topic (although see the discussion in Section 4.2 below since this possibility is not wholly uncontroversial). Thirdly, there is the possibility (iii) that the Comment is not always identical to the Focus, i.e. the Focus could be just a sub-part of the Comment as shown in (4) below. Finally, there is also the possibility of non-new foci as in (5) (viz. the discussion of second-occurrence foci; Partee, 1999).

(2) a. What do siblings do?
b. [My [SISter]$_{\text{FOC}}$]$_{\text{TOP}}$ [studies MEDicine]$_{\text{FOC}}$

(3) [A good friend of mine]$_{\text{TOP}}$ [married Britney Spears last year]$_{\text{COMMENT}}$

(4) a. When did [Aristotle Onassis]$_{\text{TOP}}$ marry Jacqueline Kennedy?
b. [He]$_{\text{TOP}}$ [married her [in 1968]$_{\text{FOC}}$]$_{\text{COMMENT}}$ (Krifka, 2007, 42–44)

(5) a. Everyone already knew that Mary only eats [vegetables]$_{\text{FOC}}$
b. If even [Paul]$_{\text{FOC}}$ knew that Mary only eats [vegetables]$_{\text{SOF}}$, then he should have suggested a different restaurant.
(Partee, 1999, 216)

Thus, the possibility of such non-canonical alignments (e.g. non-given topics, non-new foci) must be accommodated in a model of information structure. We have chosen to follow the multi-partitioning approach espoused by Krifka which assumes both a Topic/Comment and an orthogonal Focus/Background partition in order to be able to do justice to the non-canonical as well as canonical pairings.

The characterization of Aboutness Topic that we adopt is also distinct from Vallduví's (1992) "Link", which is defined positionally as the sentence-initial topic. Further, since the Focus-Background partition is independent of the Topic-Comment partition, an Aboutness Topic can in principle be identical to a focus of an utterance (though it is unclear whether or not cases other than those in (2) exist; we will return to this point below). Generally speaking, under this approach, a sentence has only one Aboutness Topic. Sentences which lack a Topic – or perhaps more precisely, a Topic-Comment articulation – are classed as thetic (cf. Krifka, 2007, 43). We will have more to say about thetic utterances in general and about the distinction between thetic vs. topic-comment utterances in particular in Section 4 below.

The guidelines for the annotation of information structure (Götze et al., 2007), which were produced by the collaborative research cluster (SFB) 632, and which closely mirror the proposals of Krifka (2007), provide instructions for the annotation of Information Status (or 'givenness'), Topic, and Focus. Under the notion of Topic, both Aboutness Topic and Frame-setting Topic are identified. It is the former that concerns us here (see Krifka, 2007, for a more detailed discussion of frame-setting). Götze et al. (2007, 165) offer the following tests for determining the Aboutness Topic of an utterance:

(6) An NP X is the Aboutness Topic of a sentence S containing X if
a. S would be a natural continuation to the announcement
Let me tell you something about X

b. S would be a good answer to the question
What about X?

c. S could be naturally transformed into the sentence
Concerning X, S'
where S' differs from S only insofar as X has been replaced by a suitable pronoun.

Applying these diagnostics to naturally-occurring data is not without problems, as will become clear in the following sections.

3 A corpus study

The initial purpose of the corpus study was to test a hypothesis about which verbal dependents (argument or adjunct) are most frequently realized as an Aboutness Topic (AT). On the basis of a prior study, occurrences of four verbs were sampled from the DeReKo corpus. The data was filtered such that only one argument-frame per verb was considered, all occurrences instantiating a different argument frame were discarded (see Table 1; a subject-XP is taken for granted in each case and therefore not listed explicitly). After also discarding occurrences in questions, relative clauses, conditionals, titles and in the first sentence of quotations, between 135 and 167 sentence tokens per verb were included in the study.

Argument frame	Verb	Example
PP_{mit} XP_{LOC}	geraten	Er gerät [mit seiner Hose] [in die Kette].
	'to get (caught in)'	'He got his trousers caught in the chain.'
XP PP_{auf}	reagieren	Sie reagiert [überrascht] [auf den Vorschlag].
	'to react'	'She reacted surprisedly to this suggestion.'
PP_{von}	profitieren	Sie profitieren [von den Steuersenkungen].
	'to profit'	'They profit from the tax reductions.'
	herrschen	Dort herrscht Ruhe.
	'to reign'	'There reigns peace.'

Table 1: Verbs and argument frames

Two independent coders (the authors of this paper) annotated a total of 587 sentenc e tokens, using the annotation schema proposed by the SFB 632 (Götze et al., 2007). The annotation task consisted in selecting the AT from among the NPs (and deictic expressions such as *hier* 'here', *dann* 'then' etc.) contained in a sentence, or alternatively stating that a sentence has no AT. Cohen's κ wa s used as a measure of inter-rater agreement, which was calculated separately for t he annotation of each verb, as shown in Table 2.[3]

[3]Cohen's k appa is the proportion of agreement that remains after chance agreement has been fa ctored out (cf. Cohen, 1960). Inter-rater agreement calculated for whether o r not a sentence was judged to contain an aboutness topic at all is also highly var iable across the four verbs: *profitieren*: κ=-.01, *herrschen* : κ=.51, *geraten*: κ=.33, *reagieren*: κ=.25; the most dramatic change is observed in *profitieren*, where kappa indicates concordance is (slightly) below chance level. This fact strongly suggests that the annotation guidelines can be interpreted in substantially different ways.

Verb	N	Coinciding Annotations	κ
profitieren	135	109	.57
herrschen	138	102	.55
geraten	147	99	**.33**
reagieren	167	106	.22

Table 2: Inter-rater agreement

Inter-rater agreement is highly variable across the four verbs, but it never exceeds κ = .57, which in our view is much less than could be expected in a case where both annotators base their judgements on the same guidelines. On inspecting more closely the tokens on which the annotators did not agree, we could identify data that proves particularly difficult to assess. Most of the controversial cases can be grouped into one of the following categories:

- Problems in deciding whether the sentence has an Aboutness Topic at all, including cases where the status of potential topic expressions is unclear because the interaction between topic and focus (especially their overlapping) is not covered exhaustively in the guidelines (nor in the literature).
- The annotators' different interpretation of "Aboutness"; most commonly, deciding "what the sentence is about" when there is more than one expression that could plausibly serve as the Aboutness Topic: in many cases, the diagnostics sketched in (6) do not yield an answer, or their application is not straightforward.

In what follows, Section 4.1 will briefly illustrate a number of cases where the annotators did in fact agree, and Section 4.2 will address examples from the two problematic categories listed above.

4 Discussion

4.1 Agreement

Examples (7)–(8) are typical of the cases in which the annotators agreed on the AT of the sentence. (In addition to the critical data (b), we also provide some of the immediately preceding context in (a).) In both examples, a non-subject was chosen as the AT. This is probably in part due to the fact that the subject, being a non-specific indefinite, is not suitable as an AT (see Endriss, 2009; Götze et al., 2007). In addition, in terms of givenness, the referent of the non-subject is either "active" (as in (7)), or "accessible" (as in (8)), which are prototypical properties of Aboutness Topics (see Section 2) above.

(7) a. Ein besonderer Fall ist der sogenannte „Promilleweg$_i$“, der von Rothenbach Richtung Brandscheid führt.
'The so-called "promille-road", leading from Rothenbach to Brandscheid, is a special case.'

b. *Auf [dem idyllisch gelegenen Wirtschaftsweg$_i$]$_{\text{TOP}}$ herrscht nämlich emsiger* on the picturesquely situated farm road reigns actually active *Autobetrieb.*
through-traffic
'The picturesque farm road is actually busy with through-traffic.'

(8) a. „Es ist schön und lustig, aber die Produktion eines solchen Spiels ist teuer, lohnen sich denn überhaupt die Kosten?“
' "It's beautiful and funny, but producing a game like this is expensive, is the cost really worth it?"'

b. *Auf [diese Frage]$_{\text{TOP}}$ würde wohl mancher Nicht-Betriebswirt mit „Typisch* on this question would probably many non-economists with typically *BWLer“ reagieren.*
economist react
'Many non-economists would probably react to this question by saying "this is typical of economists".'

Example (9) illustrates a class of cases where annotators agreed that there is no Aboutness Topic. (9b) is the first sentence of a newspaper article, with no prior context related to it except for the heading, given in (9a). However, cases similar to this one also gave rise to non-matching annotations in our study, as example (14) in the next section shows.

(9) a. Gegen Leitschiene
'Against the guardrail'

b. *Mit ihrem Pkw geriet auf der A 14 in Höhe Ortsgebiet Koblach eine* with her car got on the A 14 in height municipal.area Koblach a *Frau (18) aus Mellau ins Schleudern.*
woman (18) from Mellau into.the skid
'A woman (18 yrs.) from Mellau got into a skid on the A 14 near the municipal area of Koblach.'

4.2 Disagreement

The examples presented in this section are representative of the numerous cases that caused difficulties. Example (10b) is representative of a large number of cases that involve two expressions, each of which could justifiably be analysed as the AT of the sentence.

(10) a. *Dazugelernt habe ich besonders im Bereich der Öffentlichkeitsarbeit. Ich merkte, welche Handlung welche Reaktion auslöst und wie man gewisse Ereignisse richtig kommuniziert.*
'I learned more in the area of public relations work in particular. I noticed what sort of reaction was caused by which actions and how to communicate certain events correctly.'

b. *Von [dieser Erfahrung]$_{\text{TOP?}}$ kann [ich]$_{\text{TOP?}}$ am neuen Ort selbstverständlich* from this experience can I at.the new place evidently *profitieren*
profit
'I will clearly be able to profit from this experience at the new place.'

In many cases, one of these candidates is a non-subject that is realized in initial position. However, we do not adopt Vallduví's (1992) approach of identifying the aboutness topic positionally, as it is well known that the latter can occupy positions other than the initial position in German. The difficulty lies in deciding whether the prominent position of the PP should have priority over the fact that (i) the subject is commonly considered the default topic of a sentence and (ii) the topic of preceding sentences (in (10a), arguably the subject) is likely to be the topic of the current sentence as well ("topic chain"; see Givón, 1983).

Example (11b) is similar to (10b) in that it, too, contains two candidate expressions, but it also differs from (10) because one of these expressions (namely the subject NP) should probably bear a focus accent.

(11) a. *Diese Busspur ermöglicht die neue Buslinie, die ab 1. Juni eingerichtet wird: (...) Damit erhalten zum Beispiel die Bretzenheimer einen flotteren Anschluß nach Hechtsheim (...) Auch in die Altstadt geht's schneller.*
'This bus lane made possible the new bus route, which will operate as of June 1st: (...) The residents of Bretzenheimer will thus have a better connection to Hechtsheim (...) It will be even quicker to get into the old town-centre too.'

b. *Außerdem profitiert [der ORN-Bus aus Nieder-Olm]*$_{\text{TOP?}}$ *von [der Spur]*$_{\text{TOP?}}$
furthermore profits the ORN-Bus from Nieder-Olm from the lane
'The ORN-Bus from Nieder-Olm will also benefit from the lane.'

Note that the two possible choices of AT in (11) correspond to different discourse strategies: analysing *Spur* as the AT yields topic continuity (cf. Givón, 1983) as *Spur* is arguably the topic of (many of) the preceding sentences. On the other hand, choosing the subject-NP as the AT entails a topic switch.[4]

Turning now to examples (13b) and (14b), the annotators disagreed here on whether they were dealing with a topic-comment structure or rather with a topicless/thetic sentence. At the heart of the disagreement about these examples lies the question of precisely how the two orthogonal IS-partitions assumed here (Topic-Comment vs. Focus-Background) interact with one another, and in particular, how topic and focus may overlap. Various authors (e.g. Krifka, 1992; Steedman, 2000) suggest that both the topic (theme) and the comment (rheme) section of an utterance have their own focus-background structure. To our knowledge, the only cases discussed in which topic and focus overlap are cases of so-called contrastive topic; that is, they involve a semantic focus (marked by a rise) within the initial phrase that induces alternatives in addition to a focus later in the clause which also induces alternatives. The overall function is to indicate a discourse strategy whereby only a question that is subordinate to the (possibly implicit) question under discussion is answered. Independently of such discourse configurations, the question of the possible overlap of Topic and Focus has been less explicitly spelt out. For one annotator, there is no intrinsic problem with a complete

[4]In the terminology of the the Prague School (Daneš, 1974), these strategies correspond to a 'thematic progression' with a continuous theme and a thematic progression with derived themes, respectively. The latter decribes a configuration where there is one 'hypertheme' (i.e., a discourse topic; the bus lane, in our example), on which individual sentences elaborate. Each one of these sentence presents a theme of its own that is 'derived' in some way from the hypertheme.

overlap of (new-information) focus and AT as sketched in (12), but the other annotator tends to rule this out:

(12) Q. Who ate the apple?
A. Kim ate the apple.
[ ]FOC [ ]BACKGROUND
[ ]TOP [ ]COMMENT

If one disallows a total overlap of Topic and Focus as in (12), then the question is how the utterance should instead be analysed. One possibility which we will discuss below is that examples such as (13b) and probably also (14b) are topicless sentences. This, however, raises questions about the possible complexity of thetic utterances.

(13) a. *In Wil wird das seit Anfang Oktober gültige Rauchverbot nicht überall umgesetzt, und in gewissen Lokalen wird noch immer geraucht. Häufig wird der Gast darauf aufmerksam gemacht, dass es in seiner Verantwortung liegt, zu rauchen.*
'The smoking ban that has been in place since the beginning of October is not put into practice everywhere in Wil and people still smoke in certain bars. Frequently the customer is told that they're smoking at their own risk.'

b. *Eine andere Stimmung herrscht im [Kirchberger Restaurant Eintracht]*$_{\text{TOP?}}$*, wo das Rauchverbot strikt eingehalten wird.*
a different atmosphere reigns in.the Kirchbergian Restaurant Eintracht where the smoking.ban strictly kept is
'It's a different situation at Kirchberg's Eintracht Restaurant, where the smoking ban is strictly adhered to.'

(14) a. *Großes Bedauern über Becks Rücktritt*
'Deep regret over Beck's Resignation'

b. *Mit großem Bedauern und totaler Überraschung reagierte gestern [die Ludwigshafener SPD-Prominenz]*$_{\text{TOP?}}$ *auf den Rücktritt des Bundesvorsitzenden Kurt Beck.*
with big regret and total surprise reacted yesterday the Ludwigshafen SPD-dignitaries on the resignation of.the federal party leader Kurt Beck
'The SPD-dignitaries in Ludwigshafen reacted with deep regret and utter shock at the resignation of the party leader Kurt Beck.'

Both annotators agree that the example in (13) can be analyzed as introducing a new referent in the main clause, about which the relative clause makes a further predication. The actual information structure within the main clause itself is, however, not so evident. One annotator selected *Kirchberger Restaurant Eintracht* as the AT of the main clause, irrespective of the fact that the same phrase appears to coincide with the final focus of the main clause. The other annotator elected that there was no AT in the main clause (i.e. the introduced referent does not function as Topic until later, in the relative clause). Note that the only other potential Topic candidate, the subject NP, as a non-specific indefinite cannot normally function as an Aboutness Topic.[5] Under the latter

[5]It is worth noticing here that (13b) might be a case of i-topicalisation (Jacobs, 1997): both *eine andere Stimmung* and *Kirchberger Restaurant Eintracht* are contrasted against elements that have been previously mentioned or can be inferred from the preceding text. However, identifying (13b) as i-topicalisation does not help in deciding whether or not the sentence has an Aboutness Topic, for it is has been shown that i-topics behave differently and crucially do not necessarily involve Aboutness. See in particular Jacobs (2001); Büring (2003); Krifka (2007).

view, the main clause does not constitute a Topic-Comment utterance at all. Lacking a Topic-Comment partition is one of the defining features of thetic utterances (e. g. Lambrecht, 1994; Krifka, 2007), but classifying (13) as thetic is not without difficulties either. It is customarily said of thetic utterances that the focus spreads across the whole utterance (e. g. Lambrecht, 1994; Rosengren, 1997), and that thetic sentences in German bear a single accent on the subject (thus, the subject phonologically integrates with the predicate) (e. g. Krifka, 1984; Sasse, 1987). However, our intuition is that (13b) requires two prosodic peaks. Furthermore, a description of this utterance as event-reporting, a further characteristic of thetics, (cf. Götze et al., 2007, 163) does not seem quite correct either since, as mentioned above, the function of sentences like (13) is to introduce or present a new entity which may then later function as Aboutness Topic in the next discourse chunk.

As for example (14), a similar situation holds. One annotator chose the subject NP as topic and the other elected that the sentence had no AT. However, this example differs from (13) in that there is no contrastive element in initial position. Further, while it was clear in (13) that the main accent falls on the PP, here it could be either on the subject NP or on the final PP. If ones assumes it to fall on the subject NP, and if one assumes this to be the AT (as one annotator did), then a similar configuration to that in (13) holds. For the other annotator, who opted for a topicless analysis, the fact that the subject-NP follows the adverb *gestern* guided the decision that it is not an AT, as (14b) does not seem to be a felicitous answer to a question like "What about the Ludwigshafen SPD-dignitaries?", at least for that annotator. The sentence is discourse-initial, preceded only by a headline, and unless the sentence final NP is to be analysed as AT (an option neither annotator took), the only remaining possibility is to classify it as lacking an AT. However, as was the case with example (13b), in its natural context, sentence (14b) requires more than a single prosodic peak, thus it does not conform to the description usually given of thetic sentences.

Thus, analyzing examples like (13b) and (14b) as thetic gives rise to difficulties unless one is willing to adopt a definition of theticity which allows for a type of thetic utterance that introduces or presents an entity (rather than a situation or event). Such a definition fits in with the approach to theticity found in Lambrecht (1994, 2000) who terms this type of thetic 'presentational' (vs. 'event-reporting') and Sasse's (1987) type referred to as 'entity-central' (vs. 'event central'). Given this bifurcation of the notion of theticity, and bearing in mind the two orthogonal dimensions of IS along which sentences are analysed in the model we are assuming, one may classify thetics in general as 'all-comment' but not necessarily as 'all-focus'. The difference between entity-central thetics and event-central thetics can then be captured by recourse to their differing focus structures. Only event-central thetics involve focus spreading across the whole utterance whilst with entity-central thetics it is merely the phrase that denotes

the introduced referent that is focused (illustrated in Figure 1). A distinction between event-reporting and entity-presenting thetics was not part of the annotation guidelines at the time of the study and has now been proposed as a modification.

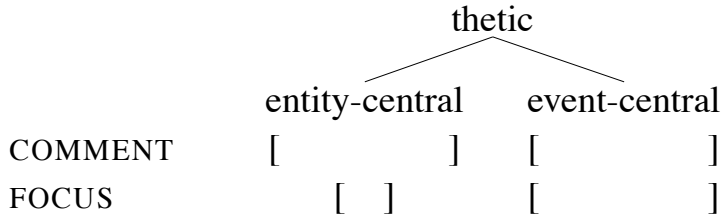


Figure 1: Analysis of different types of thetics in terms of FOCUS and COMMENT

Summing up, then, the two options for example (14) can be sketched thus (assuming that the main stress is on the subject-NP):

(15) a. [Mit großem Bedauern und totaler Überraschung reagierte gestern]$_{\text{COMMENT}}$ [[die Ludwigshafener SPD-Prominenz]$_{\text{FOC}}$]$_{\text{TOP}}$ [auf den Rücktritt des Bundesvorsitzenden Kurt Beck]$_{\text{COMMENT}}$

b. [Mit großem Bedauern und totaler Überraschung reagierte gestern [die Ludwigshafener SPD-Prominenz]$_{\text{FOC}}$ auf den Rücktritt des Bundesvorsitzenden Kurt Beck]$_{\text{COMMENT}}$

Data of the type exemplified in (13) and (14) came up frequently and the problem is thus clearly one that should be clarified in other such annotation tasks in the future. Moreover, these data show that it is necessary for annotators to state (in rough terms) the accent pattern they assumed when annotating a sentence token, as different accentuations are sometimes possible and may be indicative of different information structural partitionings.

5 Conclusion

In the present contribution, we reported on difficulties that arose from an annotation task in which we sought to operationalize the notion of Aboutness Topic. As a starting point, the annotators took the guidelines produced by a team of researchers affiliated to a collaborative research centre focusing on information structure (Götze et al., 2007). These guidelines (which, incidentally, are currently undergoing a revision phase) are undoubtedly a valuable resource and a good starting point in bringing terminological clarity to a domain of study (information structure) which is notorious for involving many conflicting definitions on the one hand but also uses of the same terminology in different senses on the other (see, e. g., Kruijff-Korbayová and Steedman, 2003). Nevertheless, once the domain of study shifts to naturally-occurring data, the concept of Aboutness Topic presents various difficulties, as thematised here.

Summing up, we hope to have alerted other researchers planning a similar enterprise to some pitfalls they may encounter and hope we can contribute to the discussion concerning issues which also have a resonance for theoretical linguistics such as (i) the

potential overlap of Aboutness Topic and focus and (ii) the correct delineation of thetic utterances and the role that presentation may play there. We remain optimistic that a careful discussion of many of the areas of contention that arose whilst conducting this study will lead to a fine-tuning of the notion of Aboutness Topic which renders it usable in future studies.

Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft within Sonderforschungsbereich 632 "Informationsstruktur" (Project A6 "Theorie und Implementation einer Analyse der Informationsstruktur im Deutschen unter besonderer Berücksichtigung der linken Satzperipherie").

References

Ben Ambridge and Adele E. Goldberg. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, 19(3):357–389, 2008.

Daniel Büring. On D-Trees, Beans, and B-Accents. *Linguistics & Philosophy*, 26(5):511–545, 2003.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Philippa Cook and Bjarne Ørsnes. Coherence with adjectives in German. In Stefan Müller, editor, *The Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, Stanford, 2010. CSLI Publications.

Philippa Cook and John Payne. Information structure and scope in German. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG06, University of Constance*, pages 124–144, Stanford, 2006. CSLI.

František Daneš. Functional sentence perspective and the organization of the text. In *Papers on Functional Sentence Perspective*, pages 106–208. Mouton, The Hague and Paris, 1974.

Kordula De Kuthy. *Discontinuous NPs in German*. CSLI Publications, Stanford, 2002.

Cornelia Endriss. *Quantificational Topics. A Scopal Treatment of Exceptional Wide Scope Phenomena*. Springer, Dordrecht, 2009.

Talmy Givón. Introduction. In Talmy Givón, editor, *Topic continuity in discourse. A quantitative cross-language study*, pages 1–41. Benjamins, Amsterdam, 1983.

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. Information structure. In Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors, *Interdisciplinary Studies on Information Structure*, volume 7 of *Working Papers of the SFB 632*, pages 147–187. Universitätsverlag, Potsdam, 2007.

Herman Hendriks and Paul Dekker. Links without locations. Information packaging and non-monotone anaphora. In Paul Dekker and Martin Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 339–358, University of Amsterdam, 1996. ILLC-Department of Philosophy.

Joachim Jacobs. I-Topikalisierung. *Linguistische Berichte*, 168:91–134, 1997.

Joachim Jacobs. The dimensions of topic-comment. *Linguistics*, 39:641–681, 2001.

Manfred Krifka. Fokus, Topik, syntaktische Struktur und semantische Interpretation. Unpublished manuscript, 1984. URL `http://amor.rz.hu-berlin.de/~h2816i3x/Publications/Krifka%201984%20Fokus.PDF`.

Manfred Krifka. A compositional semantics for multiple focus constructions. In Joachim Jacobs, editor, *Informationsstruktur und Grammatik*, pages 17–54. Westdeutscher Verlag, Opladen, 1992.

Manfred Krifka. Basic notions of information structure. In Caroline Féry, Gisbert Fanselow, and Manfred Krifka, editors, *Interdisciplinary Studies on Information Structure*, number 6 in Working Papers of the SFB 632, pages 13–56. Universitätsverlag, Potsdam, 2007.

Ivana Kruijff-Korbayová and Mark Steedman. Discourse on information structure. *Journal of Logic, Language and Information: Special Issue on Discourse and Information Structure*, 12(3):249–259, 2003.

Knud Lambrecht. *Information Structure and Sentence Form. Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, 1994.

Knud Lambrecht. When subjects behave like objects. an analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language*, 24:611–682, 2000.

Barbara H. Partee. Focus, quantification, and semantics-pragmatics issues. In Peter Bosch and Rob van der Sandt, editors, *Focus. Linguistic, cognitive, and computational perspectives*, pages 213–231. Cambridge University Press, Cambridge, 1999.

Tanya Reinhart. Pragmatics and linguistics. An analysis of sentence topics. *Philosophica*, 27:53–94, 1981.

Craige Roberts. Information structure in discourse. towards an integrated formal theory of pragmatics. In Jae Hak Yoon and Andreas Kathol, editors, *Papers in Semantics*, volume 49 of *OSU Working Papers in Linguistics*, pages 91–136. The Ohio State University Department of Linguistics, Ohio, 1996.

Inger Rosengren. The thetic/categorical distinction visited once more. *Linguistics*, 35:439–479, 1997.

Hans-Jürgen Sasse. The thetic-categorical distinction revisited. *Linguistics*, 25:511–580, 1987.

Robert Stalnaker. Assertion. In Peter Cole, editor, *Pragmatics*, number 9 in Syntax and Semantics, pages 315–332. Academic Press, New York, 1978.

Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31: 649 – 689, 2000.

Enric Vallduví. *The informational component*. Garland, New York, 1992.

Georg von der Gabelentz. Ideen zu einer vergleichenden Syntax. *Zeitschrift für Völkerpsychologie und Sprachwissenschaft*, 6:376–384, 1869.

The Lexico-Grammar of Stance: An Exploratory Analysis of Scientific Texts

Stefania Degaetano and Elke Teich
Universität des Saarlandes, Saarbrücken

Abstract
The paper reports on a corpus-based study of expressions of stance in scientific discourse. This work is part of some longer-term research on the linguistic construal of interdisciplinary scientific domains (e.g., bioinformatics or computational linguistics) compared to the disciplines from which they build mergers (e.g., biology, linguistics, computer science). We present selected analysis results on contrasts and commonalities in the use of expressions of stance across scientific disciplines using the approach of Pattern Grammar (Hunston and Francis, 2000).

1 Introduction
There is an ever growing interest in computational linguistics as well as corpus linguistics in meaning-oriented analysis of texts. While computational linguistics used to put the focus on factual content (information retrieval and extraction) in the past, there has more recently been extensive work on the extraction of opinions/sentiments from web-based documents (cf. Pang and Lee (2008); Liu (2010)). Sentiment analysis is one of the relatively new research fields that investigates opinions/sentiments from the computational linguistics point of view. There are two main approaches to extracting sentiment automatically : (1) the text classification approach, which involves building classifiers from labeled instances of texts or sentences, and (2) the lexicon-based approach, which uses dictionaries of words annotated with the word semantic orientation (polarity) (cf. Taboada et al. (forthcoming, 2)). Descriptive linguistics, in contrast, has a long-standing tradition in considering types of meaning other than the experiential (propositional content of a text; see e.g., Halliday (1985)). Accounts of the lexico-grammatical means of interpersonal meaning (stance, attitude, evaluation, appraisal, emotion and the like) can be found in most standard grammars of English, see for example Biber et al. (1999) who dedicate a whole chapter to expressions of stance. Also, there is some interesting theoretical work from different linguistic schools, e.g., Martin and White (2005) in Functional Linguistics, Reis (1999) in the framework of Generative Grammar or Hunston and Thompson (2003) in the corpus linguistic tradition, each working on different facets of interpersonal meaning and with different methodological dispositions. Hence, there is no comprehensive or uniform picture of the lexico-grammatical expression of interpersonal meaning and our understanding of it remains fairly fragmentary. Partly, this is due to the nature of the phenomena involved: first interpersonal meanings are realized in a variety of forms from lexical to structural, and second they are extremely context-dependent, both in terms of register and genre. Thus, the exact range of linguistic expressions realizing interpersonal meaning

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 57-66, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

still has to be determined and the contextual factors triggering particular interpersonal attributions and their functions in discourse have yet to be uncovered.

The present paper is situated in this field of study. We present a corpus-based analysis of one particular aspect of interpersonal meaning, *stance*, in one particular genre, *scientific research articles*. Our main interest is in the (possible) preferences of different scientific disciplines in expressing stance and we pose the following questions: How common (frequent) are expressions of stance in this genre? Which concrete stance expressions are used? How different/similar are they across disciplines? Which functions in discourse can be attributed to them? Apart from dealing with the typical problems of interpersonal analysis (identification of relevant instances, their targets and domains), we thus address explicitly the role of context (here: register/genre) in the deployment of stance expressions. Section 2 briefly describes the corpus used for the analysis. Sections 3 and 4 present the method of analysis, some first results as well as an attempt at interpretation in terms of functions in discourse. We conclude with a discussion of some theoretical, methodological and technical issues encountered in our work.

2 Corpus

The data we work on is the Darmstadt Scientific Text Corpus (DaSciTex) which contains 16.5 million words of full English scientific journal articles compiled from 23 sources covering nine scientific disciplines (Teich and Holtz, 2009). The corpus includes texts from the broader areas of humanities, science and engineering and has a three-way partition (see Figure 1): A. computer science, B. 'mixed' disciplines (B1: computational linguistics, B2: bioinformatics, B3: computer aided design/construction in mechanical engineering, B4: microelectronics/VLSI), C. 'pure' disciplines (C1: linguistics, C2: biology, C3: mechanical engineering, C4: electrical engineering).

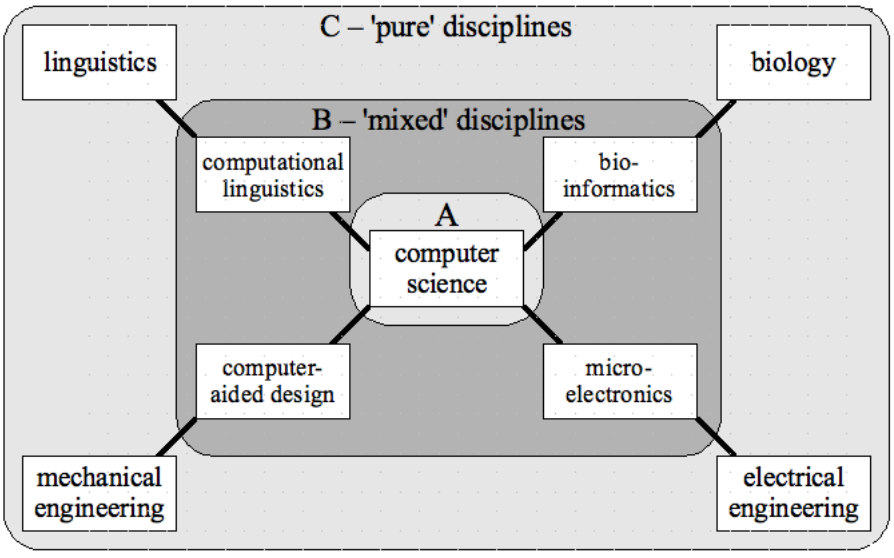


Figure 1: DaSciTex Corpus

The documents in the corpus have been enriched with meta-information (author(s), publication date and place) in accordance with TEI and the texts have been tokenized and PoS-tagged using TreeTagger (Schmid, 1994).

3 Descriptive framework and analysis

3.1 Pattern Grammar and Stance

Stance refers to one particular aspect of evaluation, which is a cover term for a speaker's or writer's attitude towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about (cf. Hunston and Thompson (2003, 5)). Conrad and Biber (2003) classify stance into three semantic classes: (1) epistemic stance, indicating the certainty, reliability, or limitations of a proposition, including comments on the source of information (*probably, according to*); (2) attitudinal stance, indicating feelings or judgments about what is said or written (*surprisingly, unfortunately*); and (3) style stance, indicating how something is said or written (*honestly, briefly*) (cf. Conrad and Biber (2003, 57), Hunston and Thompson (2003, 56)).

Evaluative lexical items are wide spread in language. They build a large and open group that is hard to quantify (cf. Hunston (2004, 157)). However, when we look in more detail at single lexical items, we can observe that they regularly appear in combination with particular words and structures that contribute to their meaning — they appear in patterns (cf. Hunston and Francis (2000)).

Although it is clearly not possible to comprehensively detect all instances of evaluative language by patterns, the pattern approach allows a fairly easy identification of particular evaluative expressions in large corpora, thus giving us one window into the exploration of evaluative meaning. Corpus-based studies of evalutive patterns may improve approaches in sentiment analysis. Especially the classification approach and its extraction pattern learning algorithms, which try to automatically identify patterns for subjective expressions, may profit from additional input (cf. Wiebe and Riloff (2003)).

One very common pattern in scientific texts is the *it v-link ADJ to* pattern (cf.Hunston and Sinclair (2003, 84)) (see example 1).

(1) *It is difficult to* tell whether or not these are possibly compounds. Certainly, many of them must at least have the option of a discontinuous analysis, since *it is possible to* adjoin a degree adverb to the P. (C1: linguistics)

This pattern begins with an introductory or anticipatory *it*, followed by a link verb, an adjective group and a to-infinitive clause. According to the local grammar of evaluation introduced by Hunston and Sinclair (Hunston and Sinclair, 2003, 84) the *thing evaluated* is located in the to-infinitive clause, whereas the *evaluative category* is realized by the adjective group. Other patterns are *it v-link ADJ that* (e.g., it is possible that), *evaluative-noun of* (e.g., importance of) and *dt most ADJ n* (e.g., the most important aspect).

3.2 Analysis

In this exploratory analysis, we focus on the pattern exemplified in (1), showing how a corpus-based analysis of evaluative patterns can contribute to further understand how evaluative meaning is expressed. For results on other evaluative patterns see Degaetano (2010).

For the detection of patterns the Corpus Query Processor (CQP) (Evert, 2005) has been used as it allows a fast corpus search by means of regular expressions in large corpora. The basis for querying is the PoS-tagged version of the DaSciTex Corpus. For the pattern under investigation here, the query in example (2) outputs any sequence of the word *it* followed by any form of the verb *be* with 0 to 3 words in between the verb and an adjective, which is followed additionally by the word *to*.

(2) `[word="it"][pos="VB.*"][]{0,3}[pos="J.*"][word="to"] within s;`

In order to be able to compare the subcorpora in DaSciTex, the adjectives appearing in this pattern have been grouped according to semantic relatedness with the help of WordNet (Miller, 1995). On this basis, four groups have been identified: (1) POSSIBILITY (e.g., *possible, impossible, feasible*), (2) IMPORTANCE (e.g., *important, necessary, relevant, vital, essential*), (3) COMPLEXITY (e.g., *diffcult, hard, simple, easy*), and (4) others (e.g., *sufficient, reasonable, useful*). The first group realizes epistemic stance, the latter three encode attitudinal stance.

4 Results and interpretation

To be able to detect possible preferences of the mixed disciplines (B subcorpora: computational linguistics, bioinformatics, computer aided design, microelectronics) in the instantiation of the pattern under study, we compare the mixed disciplines (B1-B4) to their corresponding pure disciplines (C1-C4) as well as to computer science (A). Ultimately, what we are interested in is how the mixed disciplines position themselves vis à vis the disciplines from which they build mergers: Are they closer/more similiar to computer science (A) or to their disciplines of origin (C1-C4)?[1]

Table 1 presents the results of analysis for the *it v-link ADJ to* pattern according to the four groups mentioned in Section 3.2. The most frequent group in the C subcorpora (pure disciplines) is the POSSIBILITY-group, which expresses epistemic stance (with the exception of mechanical engineering (C3) which has a slight preference for the IMPORTANCE-group). Among the B subcorpora, bioinformatics (B2) and microelectronics (B4) also make frequent use of the POSSIBILITY-group.

The comparison of the others shows that some of the engineering disciplines (computer aided design (B3), microelectronics (B4), mechanical engineering (C3)) have a preference for the IMPORTANCE-group (more than 35%), whereas computer science

[1] For a similar research question see Copestake et al. (2006), who also use patterns to identify subjectivity in the scientific domain with the difference that their work is based on recursion semantics and their focus lies on information extraction (IE).

Subcorpus	possibility		importance		complexity		others	
	F	%	F	%	F	%	F	%
A	186	32.75	71	12.50	201	35.39	110	19.37
B1	72	29.51	69	28.28	76	31.15	27	11.07
B2	144	33.64	121	28.27	103	24.07	60	14.02
B3	133	28.60	186	40.00	106	22.80	40	8.60
B4	164	38.86	150	35.55	79	18.72	29	6.87
C1	129	32.74	109	27.66	89	22.59	67	17.01
C2	75	35.38	60	28.30	53	25.00	24	11.32
C3	153	36.17	154	36.41	77	18.20	39	9.22
C4	205	35.59	149	25.87	145	25.17	77	13.37

A Computer science
B1 Computational linguistics, B2 Bioinformatics, B3 Computer Aided Design, B4 Microelectronics
C1 Linguistics, C2 Biology, C3 Mechanical Engineering, C4 Electrical Engineering

Table 1: Results of the *it v-link ADJ to* pattern

Subcorpus	p-value	significance	direction			
			possibility	importance	complexity	others
B1 - A	3.099e-07	s	–	+	–	–
B2 - A	5.979e-10	s		+	–	–
B3 - A	<2.2e-16	s		+	–	–
B4 - A	<2.2e-16	s		+	–	–
B1 - C1	0.0385	s	–		+	–
B2 - C2	0.8106	ns				
B3 - C3	0.07039	ns				
B4 - C4	5.099e-05	s		+	–	–

A Computer science
B1 Computational linguistics, B2 Bioinformatics, B3 Computer Aided Design, B4 Microelectronics
C1 Linguistics, C2 Biology, C3 Mechanical Engineering, C4 Electrical Engineering

Table 2: Comparison of subcorpora pairs: p-values of the χ^2 test

(A) and computational linguistics (B1) have a preference for the COMPLEXITY-group (more than 30%).

In order to see whether there are significant differences regarding the triples of one mixed discipline (B1-B4), its pure discipline (C1-C4) and computer science (A), chi-square values were determined for the respective combinations (see Table 2).[2]

The comparison to computer science (A) shows that the mixed disciplines make more use of the IMPORTANCE-group and less use of the COMPLEXITY-group than computer science (A).

Additionally, the comparison with the pure disciplines (C1-C4) shows that computational linguistics (B1) differs from linguistics (C1) as it makes less use of the POSSIBILITY and more use of the COMPLEXITY-group, thus being more similar to computer science (A) (due to the frequency of instances of the POSSIBILITY-group). Microelectronics (B4), instead, differs significantly from electrical engineering (C4) as it makes more use of the IMPORTANCE and less use of the COMPLEXITY-group. Looking again at B4, we can observe that it differs from computer science (A) in the same regard (more instances of IMPORTANCE, fewer instances of COMPLEXITY). Bioinformatics (B2) and computer aided design (B3) are similar to their pure disciplines, as they do not show significant differences to their pure disciplines biology (C2) and mechanical engineering (C3), respectively.

These results mainly confirm previous investigations on the mixed disciplines of DaSciTex (cf. Teich et al. (2010)) in terms of noun+verb colligations, which showed (a) a similarity of bioinformatics (B2) and computer aided design (B3) to their corresponding pure disciplines biology (C2) and mechanical engineering (C3), (b) a very pronounced distinctness of microelectronics (B4) from both computer science (A) and electrical engineering (C4) and (c) a less pronounced difference of computational linguistics (B1) to both computer science (A) and linguistics (C1).

More generally, we can deduce from the analysis results that the pattern investigated is used more to express attitudinal stance than epistemic stance (see Table 3). And within attitudinal, the IMPORTANCE-group is more common than the COMPLEXITY-group (exceptions are again computer science (A) and computational linguistics (B1) which prefer the COMPLEXITY-group).

Other interesting observations about the behaviour of evaluative patterns can be made when exploring the *thing evaluated*. In the case of the present pattern, the *thing evaluated* is a process (realized by a verbal group). Consider examples 3-6 below from the IMPORTANCE-group.

(3) *It is important to* evaluate the winglets [...] (C3)

(4) Thus, *it is important to* model the functionality (C4)

[2] s = significant, ns = not significant; '+' and '–' point to higher and lower numbers, respectively, of the category counted from the point of view of the B corpora (e.g., the difference between B1 and A is significant due to the occurrence of more instances of the IMPORTANCE-group ('+') and fewer of the other groups ('–') in B1.

	epistemic		attitudinal	
Subcorpus	F	%	F	%
A	186	32.75	382	67.25
B1	72	29.51	172	70.49
B2	144	33.64	284	66.36
B3	133	28.60	332	71.40
B4	164	38.86	258	61.14
C1	129	32.74	265	67.26
C2	75	35.38	137	64.62
C3	153	36.17	270	63.83
C4	205	35.59	371	64.41

A Computer science
B1 Computational linguistics, B2 Bioinformatics, B3 Computer Aided Design, B4 Microelectronics
C1 Linguistics, C2 Biology, C3 Mechanical Engineering, C4 Electrical Engineering

Table 3: Epistemic vs. attitudinal stance in DaSciTex

(5) *It is important to* note that the shape [...] (C3)

(6) At this point, however, *it is important to* highlight the following [...] (C4)

In examples (3) and (4) the *thing evaluated* is some domain-specific (here: engineering) activity. In example (5) it is a cognitive process (other verbs occurring are e.g., *notice, understand*) and in example (6) a semiotic process (other verbs occurring are e.g., *stress, emphasize*). Inspecting the frequency lists of these process types, in the class of cognitive verbs, the most frequently occurring verb is *note* (as in example 5), which makes up more than half of the number of occurrences in this class (see Table 4 for an overview). This remarkable difference may point to a different functional status of the pattern using *note* in the *thing evaluated*, pushing it very much in the direction of a formulaic expression with little other than stylistic meaning (and having lost the original attitudinal meaning).

5 Discussion

The attempt at more holistic interpretations of text meaning requires considering aspects of meaning other than the experiential one. Recently, interpersonal meaning (evaluation, stance, attitude etc.) has received more attention. Interpersonal text analysis poses some particular challenges, however, that are in parts still looking for better solutions. At a conceptual level, given that interpersonal meaning is highly context-dependent, very much rests upon a theory of context, i.e. one that has something to

lexical verb	F	%
bear in mind	3	1.16
consider	17	6.59
develop an understanding	1	0.39
keep in mind	3	1.16
know	4	1.55
note	152	58.91
notice	9	3.49
observe	14	5.43
predict	2	0.78
realize	10	3.88
recall	3	1.16
recognize	6	2.33
remark	5	1.94
remember	5	1.94
see	2	0.78
take into account/consideration	3	1.16
think	1	0.39
understand	18	6.98

Table 4: *important* + cognitive verb in DaSciTex

say about the relation of text and situation type, cultural domains etc. In particular, we need to take more seriously the relation of register and genre and expression of interpersonal meaning. At the methodological and technical levels, an issue to be addressed more widely is the creation of resources to facilitate analysis. Here, the availability of annotated corpora is critical (see e.g., the current activities around the American National Corpus (ANC) in terms of "crowd sourcing" for annotation (Ide et al., 2010)). Also, there is an increased need of electronic lexicons/thesauri that are enriched with interpersonal meaning categories (e.g., SentiWordNets (Esuli and Sebastiani, 2006)). Finally, provided such can be made available, specific processing work flows must be established to support linguists in conducting corpus-based analyses of stance, evaluation and the like.

This paper has presented an exploratory analysis of scientific texts in terms of stance expressions, showing how a corpus study on interpersonal meaning can reveal specific areas that deserve more intensive study. In our current work on stance in scientific writing, we investigate further the DaSciTex Corpus in terms of additional patterns observed in the existing literature on stance to find more evidence of the tendencies

of cross-disciplinary variation detected so far. In particular, we will explore in more depth the constraints between *evaluative category* and *thing evaluated* for their potentially discriminatory effects between scientific disciplines. Furthermore, knowing more about how evaluative patterns are constructed in terms of lexico-grammar, the *evaluative category* and the *thing evaluated* could be automatically identified and the value of the *evaluative category* could be automatically attributed to the *thing evaluated*. Therefore, evaluative patterns may be used to improve already existing approaches in sentiment analysis regarding the identification and the classification of evaluative language. Finally, the investigation of expressions of stance in the scientific domain and the cross-disciplinary variation may contribute to a better understanding of the expression of interpersonal meaning and genre/register variation.

References

Douglas Biber, Stig Johansson, and Geoffrey Leech. *Longman Grammar of Spoken and Written English*. Longman, Harlow, 1999.

Susan Conrad and Douglas Biber. Adverbial marking of stance in speech and writing. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text, Authorial Stance and the Construction of Discourse*, pages 56–73. Oxford University Press Inc., New York, 2003.

Ann Copestake, Peter Corbett, Peter Murray-Rust, CJ Rupp, Advaith Siddharthan, Simone Teufel, and Ben Waldron. An architecture for language processing for scientific texts. In *Proceedings of the UK e-Science Programme All Hands Meeting (AHM)*, Nottingham, UK, 2006.

Stefania Degaetano. Evaluation in academic research articles across scientific disciplines. Master's thesis, Technische Universität Darmstadt, 2010.

Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Language Resources and Evaluation (LREC)*, Genoa, May 2006.

Stefan Evert. *The CQP Query Language Tutorial*. IMS Stuttgart, 2005. CWB version 2.2.b90.

M.A.K. Halliday. *An Introduction to Functional Grammar*. Arnold, London, 1985.

Susan Hunston. Counting the uncountable: problems of identifying evaluation in a text and in a corpus. In Alan Partington, John Morley, and Louann Haarman, editors, *Corpora and Discourse*, pages 157–188. Peter Lang, 2004.

Susan Hunston and Gill Francis. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Studies in Corpus Linguistics. John Benjamins Publishing, Amsterdam/Philadelphia, 2000.

Susan Hunston and John Sinclair. A local grammar of evaluation. In *Evaluation in Text, Authorial Stance and the Construction of Discourse*, pages 74–101. Oxford University Press Inc., Oxford, 2003.

Susan Hunston and Geoff Thompson, editors. *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford University Press, Oxford, 2003.

Nancy Ide, Keith Suderman, and Brian Simms. Anc2go: A web application for customized corpus creation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, Malta, May 2010.

Bing Liu. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Goshen, Connecticut, USA, 2 edition, 2010.

Jim R. Martin and Peter R.R. White. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, London & New York, 2005.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38:39–41, 1995.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:Nos. 1–2:1–135, 2008.

Marga Reis. On sentence types in German: An enquiry into the relationship between grammar and pragmatics. *Interdisciplinary Journal for Germanic Linguistics and Semiotics Analysis*, 4:195 – 236, 1999.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, September 1994.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, forthcoming.

Elke Teich and Mônica Holtz. Scientific registers in contact. An exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics*, 14(4):524–548, 2009.

Elke Teich, Stefania Degaetano, Mônica Holtz, and Tatsiana Markovic. Creating meaning differences through colligation. Linguistic variation in scientific texts. In *European Systemic Functional Linguistics Conference and Workshop (ESFLCW)*, Koper, Slovenia, July 2010.

Janyce Wiebe and Ellen Riloff. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sapporo, Japan, July 2003.

Suggestions in British and American English: A Corpus-Linguistic Study

Ilka Flöck
Carl von Ossietzky Universität Oldenburg

Abstract

This article examines the surface realisations of the speech act of suggesting in two national varieties of English, British and American English. More specifically, it analyses and compares the head act realisations of the speech act and its internal and external devices of modification across the two speaker groups. The study is based on corpus data retrieved in automated searches from two corpora of English. The analysis reveals that despite general similarities, the two data sets differ in frequency distribution of head acts and their modification. Furthermore, the results show that the surface realisations used to encode suggestions are functionally ambiguous in that they can also be used to realise other illocutions, such as requests or orders. The paper therefore calls for the inclusion of the hearer perspective in pragmatics research to fathom out how hearers are able to infer speaker meaning. Gaining knowledge about how intention is identified will also help to improve inter-rater reliability in coding data and annotating corpora for pragmatic units.

1 Introduction

The basic insight that speech is action has become the foundation for one of the most influential theoretic frameworks in pragmatics. Speech act theory originated in the rejection of the idea that language can be described solely on the basis of formal semantics. Even with a purely philosophical starting point, speech act theory was able to trigger enormous amounts of empirical research. For many speech acts, linguistic manifestations have been established and compared across cultures. While most illocutions seem to be to be universal, their linguistic manifestations might differ sharply across cultures. Differences in realisation form (such as diverging levels of directness) can have the potential to lead to difficulties in intercultural communication. But this is not only true for communication between speakers of different languages. It has also been found that speech acts may be realised differently in national or even subnational varieties of one language. Language users usually are unaware of such intralingual differences and often attribute pragmatic variation across varieties to character flaws in the individual speaker.

The present study raises and tries to answer the question if there are similarities or differences in the realisation of the speech act of suggesting in two national varieties of the English language: British English (BrE) and American English (AmE). While other speech acts, most prominently requests, compliments and compliment responses have been studied extensively by many researchers worldwide, little is known about how suggestions are realised in English. The only studies concerned with this speech act are situated in educational linguistics that use suggestions as a diagnostic means to investigate learners' pragmatic competence. In order to do so, these studies have predominantly made use of experimental methods or recordings of natural conversations in institutional contexts. Consequently, they cannot provide any information about how native speakers of English make suggestions in naturally

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 67-81, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

occurring casual conversations. In contrast to many studies on the realisation forms of speech acts, the present study does not make use of experimental data such as questionnaire material. The material analysed comes from two language corpora, the British component to the *International Corpus of English* and the *Santa Barbara Corpus of Spoken American English*. With their vast and growing amount of language material, corpora equip researchers with a valuable tool to study speech acts in large populations across language varieties.

2 Suggestions in English

2.1 Defining the function of the speech act

In one of the most influential classifications of illocutionary acts (Searle 1976) suggestions are defined as directive speech acts since they are attempts by the speaker to "get the hearer to do something" (Searle, 1979: 12). Searle claims that the illocutionary point can be realised with varying illocutionary forces, as "modest 'attempts' as when I (...) suggest that you do it, or they may be fierce attempts as when I insist that you do it" (Searle, 1979: 13). Suggestions are thus defined to be milder attempts to get a hearer to do what the speaker wants than other directive speech acts, such as requests or orders.

Defining suggestions as directive speech acts only has, however, triggered criticism from other researchers working in the speech act theoretic paradigm. Hancher (1979) argues that some speech acts have both a commissive and a directive illocutionary point and are thus hybrid speech acts that belong to more than one of illocutionary type as defined by Searle. He gives the example of invitations which Searle categorises as directives and claims that an invitation is successful not only on the grounds that the hearer appears at the event in question (and therefore complies with the action desired by the speaker). He argues that it is also necessary for the speaker to receive the person invited as a guest. In issuing an invitation, it is thus both the hearer and the speaker who have to fulfil a future action. Invitations are therefore "hybrid speech acts that combine directive with commissive illocutionary force" (Hancher, 1979: 6). Although he does not explicitly mention suggestions to belong to this hybrid category, it is easily conceivable that they also have a commissive directive illocutionary point (cf. Adolphs, 2008: 45). In suggestions, speaker**s** can include themselves in the action proposed to the hearer, as seen in Example (1).

(1) SETH: Well, I mean -- we could put a floor r- .. floor register right .. along here (SBC 071)

This functional bipolarity of suggestions is also recognised in studies that are concerned with the forms and functions of speech acts but were not conducted within the paradigm of speech act theory. In their discourse oriented interactional grammar, Edmondson and House (1981) state that suggestions can include the speaker in a future joint action. The authors therefore distinguish between suggestions that exclude the speaker ("suggests-for-you") and suggestions that include the speaker ("suggests-for-us"). In a similar vein, Tsui distinguishes suggestions from requests in that a "request for action prospects only addressee action" (Tsui, 1994: 100).

In an alternative approach to classifying speech acts, Fraser (1974: 149) defines suggestions as speech acts in which the speaker "indicates his desire for the hearer to consider the merits of the state of affairs expressed by the proposition". Subtypes of the speech act class of suggesting include suggesting proper, imploring, recommending and advising. Suggestions are defined as speech acts that are always in the interest of the hearer. Fraser also claims that the action anticipated in requests and advice is not of the same kind. The action that the hearer is supposed to fulfil is rather a cognitive process than a physical action. The speaker wants the hearer "to consider the merits" of the action proposed. In Fraser's definition of suggestions the hearer has the option to conclude that the action proposed is not convergent with her own intentions.

The same position is proposed in Hindelang's (1978) comprehensive classification of directive speech acts. He defines suggestions as non-binding directives which do not put the hearer under the obligation to comply with the action proposed by the speaker. He subcategorises suggestions further into problem solving suggestions and proposals. While problem-solving suggestions are always task-related, proposals are not associated with a practical problem and come closest to suggestions as dealt with in this study.

The definition adopted for the speech act of suggesting in the present study is a combination of definitions reported on above. A speech act is understood as a suggestion when the following conditions apply:

– The speaker (S) wants the hearer (H) to consider the action proposed.
– S and H know that H is not obliged to carry out the action proposed by S.
– S believes that the suggestion is in the interest of H.
– S may or may not include herself in the proposed action.

2.2 Defining the form of the speech act

Insights about the linguistic forms that suggestions can take come from sources that are different in aim and methodological setup. Empirical investigations on suggestions have their origins predominantly in the field of interlanguage pragmatics in the English as a Foreign Language (EFL) contexts. In most of those studies it is not the speech act itself that is of interest for the authors but the learners' pragmatic competence. Suggestions only serve as a diagnostic means of identifying the degree of learner competence. Due to their research questions the vast majority of these studies made use of experimental data to measure learners' improvements after pragmatic construction (e.g. Martínez Flor, 2004) or compare learners' and native speakers' pragmatic competence (e.g. Rintell, 1979; Banerjee and Carrell, 1988; Koike 1994; 1996). Considering their didactic starting point, it is not surprising that in neither of these studies the speech act of suggesting has been studied systematically. The studies do, however, provide information about the linguistic forms that the speech act of suggesting may take.

A different source for information about linguistic surface structures of speech acts are communicative grammars that investigate both written and spoken language (e.g.

Edmondson and House, 1981; Leech and Svartvik, 1994; Carter and McCarthy, 2006). Overall, 60 realisation forms of suggestions were found in the literature (cf. Table 1).

Linguistic form	Source
Can't we	Edmondson and House, 1981
Can't you	Koike, 1994; Carter and McCarthy, 2006; Adolphs 2008
How about	Leech and Svartvik, 1994; Koike 1994; Carter and McCarthy, 2007; Adolphs, 2004
I (would) suggest	Edmondson and House, 1981; Leech and Svartvik, 1994; Martínez Flor, 2004; Adolphs, 2008
Let's	Sadock, 1974; Edmondson and House, 1981; Koike, 1994
Shall/ should we	Edmondson and House, 1981; Leech and Svartvik, 1994; Carter and McCarthy, 2006; Adolphs, 2008
We can/ could	Edmondson and House, 1981; Koike, 1994; Martínez Flor, 2004; Carter and McCarthy, 2006
What about	Leech and Svartvik, 1994; Carter and McCarthy, 2006; Adolphs 2008
Why don't we/you	Leech and Svartvik, 1994; Koike, 1994; Martínez Flor, 2004; Carter and McCarthy, 2006; Adolphs, 2008
You/we can/could	Leech and Svartvik, 1994; Koike, 1994; Martínez Flor, 2004; Carter and McCarthy, 2006

Table 1: Overview of the most frequently cited linguistic forms realising suggestions.

It needs to be acknowledged that the different sources for linguistic forms differ in aims and approaches to obtaining the linguistic surface manifestations. While most of the didactic studies on suggestions are based on experimental data, the communicative grammars make use of 'field' data (cf. Jucker, 2009). While studies investigating speech acts usually take a function-to-form approach, corpus linguistic investigations make a form-to-function approach necessary. The differences between the approaches and the implications for studying pragmatic variables will be outlined in the following chapter.

3 Methodology

3.1 Speech acts and corpora

While corpora have predominantly been used in research on lexicography and grammar, they are gaining more and more importance in other linguistic disciplines such as pragmatics (cf. McCarthy and Carter, 2004). The use of language corpora in pragmatics, and more specifically in the investigation of speech acts, is, however, problematic to some degree. While the starting point in corpus linguistics is always a linguistic form that is to be searched for in a corpus, pragmatics often takes a functional perspective. Language functions, however, do not lend themselves to searches in language corpora per se. While many corpora available today are tagged

for parts of speech or even parsed for sentence structures, there are no corpora available that are tagged for speech acts. Consequently, in their study on compliments in the *British National Corpus* (BNC), Jucker et al. claim that speech acts "are not readily amenable to corpus-linguistic investigations" (Jucker et al., 2008: 273). The authors explain that speech acts are defined by their illocutionary force or their perlocutionary effect, neither of which can be searched for directly in a corpus. Speech acts can therefore only be searched for in language corpora when they appear in routinised forms or in regular combination with illocutionary force indicating devices. In the case of compliments, linguistic forms or formulae had already been established (cf. Manes and Wolfson, 1981) enabling Jucker et al. to trace the speech act in the BNC.

There are a number of speech acts and discourse features whose forms have either been investigated thoroughly in past research (as for compliments) or occur in a routinised form. In a contrastive study on thanking, Jautz (2008: 147) observes that expressions of gratitude are "highly ritualised formulae" that can be searched for easily in a corpus. With a list of forms expressing gratitude established in earlier research on thanking, Jautz conducts word searches in the BNC and the *Wellington Spoken Corpus* (WSC) and compares the head acts and modifiers used in radio phone-ins in the BrE and New Zealand English. In a study on listenership in everyday BrE and AmE discourse, McCarthy (2002) traces non-minimal response tokens in the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE). The selection of search items from a list of the 2000 most frequent lexical items in both corpora was based on the forms of response tokens established in earlier research. In a similar study, O'Keeffe and Adolphs (2008) rely on the forms established for response tokens in previous research for their corpus searches in the CANCODE and the *Limerick Corpus of Irish English* (LCIE). But even if formulae are available that can be used as search strings in the corpus query, problems of precision and recall may emerge. Jucker et al. (2008) note that searches for relevant patterns may retrieve large numbers of hits that are identical in structure but not in function (low precision). These extracts then have to be filtered manually for function and excluded if they do not realise the functional unit in question. This procedure, however, is only possible until the number of hits exceeds what is possible to analyse manually. Problems of recall occur both on the level of word queries and queries for syntactic strings. Word queries might not have a complete recall since typing errors or different spelling conventions (especially for minimal response tokens such as *uhunh*) prevent the items in the corpus from being found. Queries for syntactic patterns are even more prone to incomplete recall since it is impossible to account for all possible sequences when tagging or parsing a corpus (cf. Jucker et al., 2008).

An approach that tries to overcome these methodological problems in using corpora for pragmatic research is Kohnen's (2008) study on directives in the history of English. The author stresses that with automated searches alone, it is impossible to access all manifestations of a particular speech act in a past period. This argument is also valid when investigating a speech act synchronically that has not been studied extensively. Kohnen puts forward that even in those cases where formulae have been

established in earlier research, it can never be ruled out that realisation forms are not accounted for by the corpus searches. He argues that studies relying on forms established in earlier research cannot "exclude the possibility that some other manifestations of the speech act are hidden somewhere in the corpus" (Kohnen, 2008: 295). Consequently, the author starts from a different point of departure. His genre-based micro-analytic bottom-up approach comprises first a manual search of a corpus limited to one genre. Since the task is reported to be "extremely labour-intensive" (Kohnen, 2008: 296), the corpus for the initial selection of realisation forms must be limited in size. In a second step, this procedure is repeated for corpora of different genres before finally testing the manifestations established this way by searches in larger corpora of mixed genres. Since the initial corpus needs to be relatively small, Kohnen's approach cannot guarantee either that all possible realisation forms of a speech act can be found in corpora.

For the analysis of suggestions in the present study, a top-down approach was chosen due to time restrictions and the fact that the forms of suggestions have already been established in the literature. It is assumed that these realisation forms will represent the high frequency manifestations of suggestions. Indirect and low frequency realisation forms cannot, however, be guaranteed to be accounted for by the present study.

3.2 Data collection and coding

The data for the present study were collected using automated searches of two corpora representing national varieties of English, the *Santa Barbara Corpus of Spoken American English* (SBCSAE) and the British component to the *International Corpus of English* (ICE-GB) (cf. Section 3.3 for information on the subcorpora established for the present study). The realisation forms reported on in the literature (cf. Section 2.2) were used as search tokens. While for searching ICE-GB, the utility program *ICECUP* 3 was used, the SBCSAE was searched with the concordance sampler of *WordSmith Tools* 5.0.

All the hits were then analysed for function, excluding all tokens which were not identified as suggestions on the basis of the definitions illustrated in Section 2.1. The data sets gathered that way comprise 233 tokens of suggestions (117 tokens in the BrE data set, 116 tokens in the AmE data set). The coding scheme adopted in the present study is based on Blum-Kulka et al.'s (1989a) coding system developed for the comparison of two speech acts, requests and apologies, across different languages. It differentiates between the head act, internal modification and external modification. The head act is defined as "the minimal unit which can realize a request" (Blum-Kulka et al. 1989a: 275). Internal modification includes syntactic downgraders which "modify the head act internally by mitigating the impositive force" (Blum-Kulka et al. 1989a: 281) of the speech act by means of syntactic choices", lexical and phrasal downgraders, which in analogy to syntactic downgraders modify the head act internally by means of lexical or phrasal choices and upgraders. The latter are defined as "elements whose function it is to increase the impact of the request" (Blum-Kulka et al., 1989a: 285). Table 2 gives an overview of

the most frequently occurring modifiers found in the data.

Due to the abstract definitions of head acts and modification, the coding system as proposed by Blum-Kulka et al. (1989a) can easily be applied to the analysis of illocutions other than requests and apologies.

	Modifier	Example
Mitigating	Conditional	*You **could** help it by...*
	Pseudo-cleft	***What I recommend you do Tony** is ...*
	Concluder	*so..., then..., well..., well then...*
	Understater	*a bit, to begin with, for the moment*
	Hedge	*sort of, something, like, somehow*
	Subjectivizer	*I think, I mean, I would say*
	Downtoner	*just, perhaps, at least, maybe, probably*
	Grounder	*You should go. **They keep saying where's Louisa***
	Specification	***I mean that would have to be moved anyway***
	Antecedence present	*We can get that out **if you want***
Aggravating	Repetition	B: *You should stay.* [...] B: ***You should stay.***
	Contradicting hearer	*We can get that out. **But I d I don't think**...*
	Consequences	*Otherwise you gotta come back and put the coil in*
	Intensifier	*I'd **highly** recommend...*
	Negative interrogative	*Well **can't you just** ring the the company direct?*

Table 2: Overview of the most frequently occurring modifiers in the data sets.

3.3 Corpora used in the present study

For the present study, subcorpora from the British component of the *International Corpus of English* (ICE-GB) and the *Santa Barbara Corpus of Spoken American English* (SBCSAE) were chosen to examine and compare the realisation strategies used for the speech act of suggesting in BrE and AmE. ICE-GB includes 200 written and 300 spoken (and transcribed) text samples of about 2,000 words each, adding up to a total of around 1,000,000 words. In contrast to ICE-GB, the SBCSAE predominantly includes transcripts of casual conversations. The four parts of the corpus amount to approximately 249,000 words. The SBCSAE was sampled with the aim of providing a source of data for researchers "interested in the nature of spoken American English" (Chafe et al. 1991: 65) in descriptive, theoretical or pedagogical contexts. The language material in both corpora was sampled in the 1990s.

From both corpora smaller subcorpora were selected that are highly comparable in terms of linguistic genre included and speaker demographics. The ICE-GB subcorpus used in the present study consists only of 100 transcripts of direct and telephone conversation. From the SBCSAE all scripted material such as lectures, sermons and transcripts from guided tourist tours were excluded from analysis, leaving a total of 50 transcripts. Since the individual samples of the SBCSAE are much longer than the text samples in ICE, each subcorpus has approximately the size of 200,000 words.

4 Results

The results reveal that there are only mild differences between the two national varieties of English. The analysis shows that both varieties use very similar head act super- and substrategies with approximately the same frequency. In the vast majority of cases, suggestions were realised by modal head acts (55.6% in the BrE, 61.2% in the AmE data set) or specific formulae (35.9% in the BrE, 37.9% in the AmE data set), which are syntactically fixed expressions closely associated with the speech act (such as *let's* and *why don't you*). With a proportion of 6.0% of all head acts in the BrE data (0.9% in the AmE data set) the superstrategy of performative utterances (such as *suggest* and *recommend*) was used only infrequently. The utterances *I'd if I were you* and *You'd (..) better* only occurred in the BrE data and accounted for only 2.6% of head acts. The differences in distribution in the two groups did not reveal to be of statistical significance.

Within the head act superstrategies, only a few substrategies were used with high frequency. The five most frequent substrategies in the data sets are presented in Figure 1 below. The remaining 13 substrategies in the BrE data set (11 in the AmE data) occurred with very low frequencies (n = fewer than four hits) in both data sets.

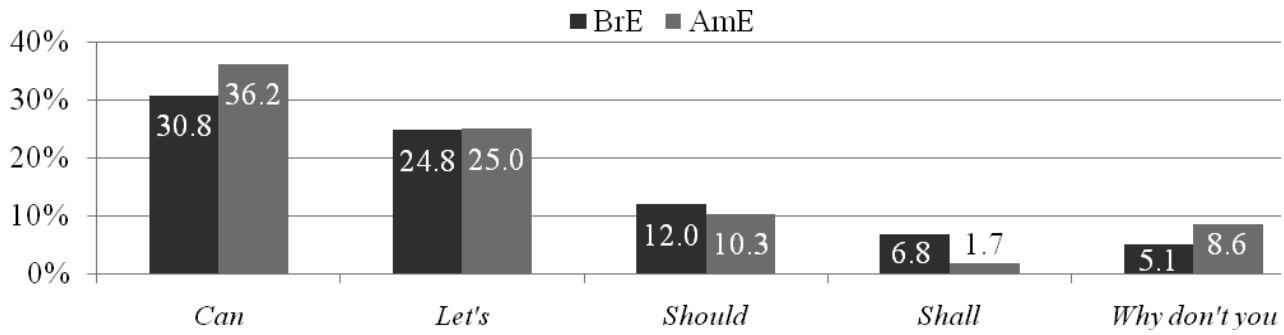


Figure 1: Distribution of the most frequent head act substrategies.

In the most frequent head act superstrategy of modals, some differences emerge in the BrE and AmE data sets. While both groups employed modal verbs of possibility most often (BrE 55.4%, AmE 59.2% of all modal head acts), British speakers showed a tendency to use modals of obligation more often than their American counterparts (44.6% in the BrE group, 36.6% in the AmE group). The usage of modals of obligation can be interpreted to be a more direct strategy in realising suggestions since the speaker imposes on the hearer's freedom of action more strongly.

The differences between the two groups become more pronounced when analysing the frequency and kind of modification used. Since modifiers can serve two different functions – downgrading or mitigating and upgrading or aggravating the head act – they were clustered for function in the present analysis. With 1.6 modifiers per head act, the British group overall used more modifiers than the American group (1.4 modifiers per head act). In both data sets, modifiers with a mitigating function were used in the vast majority of cases (cf. Figure 2). The British group, however, displayed a stronger preference for upgraders than the American group. The difference in frequency distribution was found to be statistically significant in

ANOVA testing (F (1.231) = 4.926, $p < 0.05$). The higher use of upgraders in the BrE group can be traced back to the more frequent occurrence of the negative interrogative structure. While studies on suggestions (Koike, 1994; 1996) define this structure as an aggravating device, it has been defined as a mitigating modifier in empirical investigations of requests in BrE and AmE (cf. Breuer and Geluykens, 2007). In their questionnaire-based study of requests, the authors find the negative interrogative structure exclusively in BrE requests. It is therefore questionable if the negative interrogative serves an aggravating function in suggestions while it serves as a mitigating device in requests. Since numbers of occurrences for negative interrogatives in the present study were very low, it would be necessary to explore the function of this form in BrE in a larger sample.

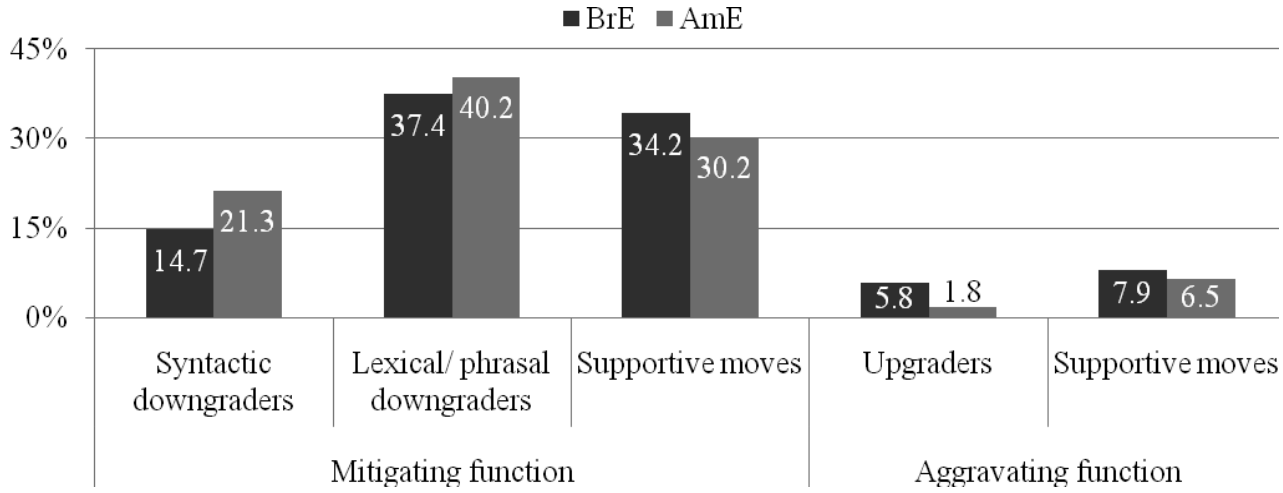


Figure 2: Distribution of modifiers in the data sets.

When analysing the distribution of modifiers among the different head act strategies, it becomes apparent that some head act strategies are more heavily modified than others at statistically significant levels. The choice of head act strategy is therefore an determining factor in the frequency of modifiers used. The most striking differences in distribution can be found in modal and specific formulae head act strategies. Even granted the fact that due to their rigid structure specific formulae are incompatible with syntactic downgrading, they still are combined with fewer lexical and phrasal modifiers and supportive moves relative to the other head act strategies. In this context, it is important to keep in mind that suggestions are speech acts which are uttered in the interest of the hearer. With this characteristic, they are thought to be less face-threatening than other speech acts with a directive force. In their programmatic work on verbal politeness, Brown and Levinson (1987) claim that speech acts in the interest of the hearer do not need to receive redressive action (e.g. mitigating modifiers) at all. With the low levels of modification, it can therefore be assumed that specific formulae are so strongly associated with the speech act of suggesting that they do not need to be softened in all instances. Since modal head acts are not only used to realise suggestions, speakers seem to find it necessary to combine them with mitigating devices to signal that their utterance should be understood as a suggestion and not as a more binding directive. This interpretation is

supported by the fact that the numbers of modifiers among requests in the same national varieties as examined in the present study are higher than for suggestions. For BrE requests, Breuer and Geluykens (2007) find a mean of 2.62 modifiers with a mitigating function per head act. A mean number of 1.9 mitigating modifiers was found for the American group. With 1.4 mitigating modifiers per head act in the British and 1.3 mitigating modifiers in the American group, the levels of modifiers with a mitigating function are therefore much lower for suggestions.

5 Discussion

5.1 Comparability of results: Implications of genre and method

The present study uses naturally occurring casual conversation as a basis for establishing realisation forms for the speech act of suggesting in two varieties of English. With this approach, it differs sharply from all other studies on suggestions. As pointed out in Section 2.2, most of these studies have made use of experimental data. There are, however, studies that have based their findings on recordings of naturally occurring talk (cf. e.g. Bardovi-Harlig and Hartford, 1990). But these studies also differ from the present paper in regard to the genre included. While the present study includes only recordings of casual conversation, Bardovi-Harlig and Hartford's (1990) study is set in an academic advisory context. Most of the studies employing questionnaires also use the context of academic advising more or less explicitly. Therefore, all of the studies on suggesting differ from the present study in the method of data elicitation, the genre included or both.

It is, therefore, difficult to compare the results of the present study to the findings of previous research. Since many studies differ from the present one in more than one variable, it is even more difficult to distinguish which differences can be accounted for in terms of methodology used or in terms of genre included. When comparing the number of realisation forms found in the literature with the strategies found in the present study, it becomes apparent that only a fraction of all these forms was employed by the speakers of the two data sets. While 60 realisation forms have been identified by various authors, only 18 strategies were employed in the BrE and 16 in the AmE data set. To discover if genre and the context of academic advisory session had an influence on the selection of realisation strategies by other researchers' informants, the realisation forms established were searched for in a specialised corpus to see if they are genre specific. The corpus selected for this attempt is *The Michigan Corpus of Academic Spoken English* (MICASE) which consists of spoken academic English only. This genre is further divided into different subgenres, such as lectures, meetings, office hours and advisory sessions. While those forms that were used frequently as realisation strategies for suggestions in the present data sets were also used frequently in MICASE, strategies that were not used at all in the present study could be found in academic contexts. Table 3 gives an overview of selected strategies which occurred in academic contexts only. The frequencies with which these strategies occur are, however, relatively low. It is unfortunate that Martínez Flor (2004) and other authors do not provide information of how frequently the realisation

strategies they identified were used among their informants. A quantitative comparison is therefore impossible.

Linguistic form	MICASE	ICE-GB	SBCSAE
Have you thought about...	3	0	0
It might be better to...	3	0	0
It would be a good idea to...	6	0	0
One thing you can do is...	1	0	0
The thing to do is...	2	0	0
There are a number of options that you...	6	0	0

Table 3: Absolute frequencies of strategies used in academic advising contexts only.

The influence of methodology on the use of realisation strategies is less transparent. Questionnaire data has been reported to elicit rather the culturally expected forms of speech acts than the forms actually used (cf. Beebe and Cummings, 1996). In a comparison of DCTs and naturally occurring talk, Golato (2003) finds differences in the realisation of compliment responses in data elicited by DCTs and naturally occurring data. She reports that the use of the appreciation token *thank you* in combination with other strategies or on its own is much higher in the DCT data than in the natural data. The author argues that this finding can be attributed to social expectations. When filling in a written questionnaire, many informants provide the response they think is socially expected rather than writing what they would actually say in natural conversation. She also finds that both methods of data collection overall produced the same strategies of responding to compliments. The DCT data, however, differ from naturally occurring data in that participants produced more combinations of strategies. The responses were generally longer in DCTs. Golato explains this finding by the absence of an interlocutor in questionnaire settings and argues that speakers might self-select if no response comes from the interlocutor. This self-selection then causes speakers to produce more turns and therefore longer responses in questionnaires where no interlocutor is present. In a similar study on compliment responses in Mandarin, Yuan (2001) finds similar differences in length and number of turns. The author also accounts the greater length in DCT responses to the missing interaction between interlocutors.

Since none of the studies dealing with suggestions provide information about the frequencies of individual realisation strategies, it is not possible to detect if this also was the case for the studies in question. It is, however, easily perceivable that methodological differences also had an impact on the choice and kinds of realisation forms that have been established for suggestions so far.

5.2 Identifying functional units in natural conversation

Identifying functional units without standardized and distinctive surface manifestations in conversation – or more generally, in all kinds of non-elicited language material – proves to be problematic. The surface structures found for suggestions in the present study can also be used to realize other illocutions such as

requests or offers. This functional ambiguity of realization forms is even reflected terminologically in the head act realisation strategy "suggestory formula" in Blum-Kulka et al. (1989b). The authors state that requests may be realised by utterances "which contain a suggestion" (Blum-Kulka et al., 1989b: 18) to carry out the action longed for by the speaker. Suggestory formulae in requests are defined as strategies of conventional indirectness by the authors. Although the illocutionary point is not retrievable directly from the linguistic form, it is the conventionalized character of such utterances that makes them interpretable as requests by the hearer. Searle (1975: 76) suggests that some linguistic forms become "conventionally established as the standard idiomatic forms for indirect speech acts". While they keep their literal meanings, "they will acquire conventional uses as, e.g. polite forms for requests" (Searle, 1975: 76). Trosborg (1995: 201) specifies this function of suggestory formulae in that the strategy is employed when requesters test "the hearer's cooperativeness in general by inquiring whether any conditions exist that might prevent the hearer from carrying out the action specified by the proposition". The speaker is therefore able to make her request more tentative and "plays down his/her interest as a beneficiary of the action" (Trosborg, 1995: 201). The defining property of suggestions, i.e. the action being in the interest of the hearer, is therefore transferred to the speech act of requesting when using suggestory formulae. In employing this strategy, the speaker pretends that the action might also be in the interest of the hearer while in reality it is in the sole interest of the speaker. The use of suggestory formulae can therefore be understood as a strategy of conventional indirectness in requests, in suggestions proper, however, they are a means of literally and directly realizing the speech act.

The terminological confusion of illocutions is symptomatic for the focus on the speaker perspective in speech act theory or the empirical study of speech acts. Whereas the speaker perspective has been explored extensively for speech acts such as requests, apologies and compliments, the hearer perspective – or more specifically the question of intention recognition – has received noticeably less attention in speech act research. This lack of knowledge about how hearers are able to infer speaker meaning is not only regrettable from a theoretic point of view but also has implications for researchers tracing functional units in conversation. Until we know which factors (such as linguistic surface manifestation, context or cotext) are involved in intention recognition, we have to accept that identifying functional units in non-elicited language material is a more or less subjective matter which can only be partially remedied by including several researchers in the coding process and comparing inter-rater reliability.

One of the few noticeable exceptions from the lack of research into the hearer perspective is Herbert Clark's work on language perception. While Clark has investigated the hearer perspective for some illocutions (e.g. Clark and Lucy, 1975), or how common ground between interlocutors is established (e.g. Clark and Brennan, 1991), his research is not aimed at answering the questions whether (or how) very similar illocutions such as requesting, suggesting or advising are perceived differently by hearers in natural conversations. An interesting starting point for this

kind of investigation is offered by Thomas Holtgraves' (2007) studies in automatic intention recognition. Holtgraves finds that informants are able to activate metapragmatic knowledge when being confronted with samples of speech acts in authentic conversations. When manipulating these sets of speech acts linguistically and contextually, informants do not show metapragmatic activation. Unfortunately, Holtgraves uses only four linguistic variables that he systematically varies in different scenarios (switching tense/ subject, negating speech act and replacing original speech act with a different illocution). As valid as these variables may be for the activation of metapragmatic knowledge, they do not provide any exhaustive evidence as to the linguistic or contextual factors that are involved in identifying speaker meaning.

Research into how hearers are able to infer speaker meaning – or more specifically identifying different illocutions – is therefore crucial in aiming at a more objective and reliable approach to identifying functional units in non-elicited language samples.

6 Conclusion

The present study has compared realisation forms for the speech act of suggesting in corpora representing two varieties of English: British and American English. A correlation of head act strategy and modification devices showed that it is rather the most frequently used strategy than the most direct strategy that receives the highest levels of modification. This trend was observed for both varieties. Apart from modest preferences for one or the other head act or modification strategy, no major differences between the two varieties could be observed. Unlike other speech acts, suggestions might therefore not have a strong potential for intercultural misunderstanding. The different trends of realising suggestions should, however, be investigated in larger samples of conversation to confirm the present results.

The study raises, however, a more general question about different illocutions. Many of the realization forms for suggestions cannot be distinguished from the forms realizing other illocutions on the formal level. The question still remains unanswered how hearers are able to detect the speaker's meaning if linguistic forms can be used to realize more than one illocution. Given that the perception of the speaker's intention seems to play an important role in understanding how hearers comprehend discourse, it is essential to answer this question. When perception is to be investigated, methods are to be employed that are able to give insights into cognitive processes in the hearer. The pragmatic apparatus of methods needs to be supplemented with psycholinguistic methods to get a fuller understanding of how speech acts are produced, comprehended and negotiated between interlocutors. Pragmatics, therefore, needs to come together with psycholinguistics to answer the question of how illocution or speaker intention is understood by the hearer and on a more applied level with computational linguistics to discuss if and find ways of how functional units can be coded in language corpora.

References

Adolphs, Svenja (2008): *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam/ Philadelphia: Benjamins.

Banerjee, Janet & Carell, Patricia L. (1988): "Tuck in your shirt, you squid. Suggestions in ESL." In: *Language Learning* 38 (3), 313-364.

Bardovi-Harlig, Kathleen & Hartford, Beverly (1990): "Congruence in native and nonnative conversations. Status balance in the academic advisory session". In: *Language Learning* 40 (4), 467-501.

Beebe, Leslie & Cummings, Martha (1996): "Natural speech act data versus written questionnaire data: How data collection method affects speech act performance". In: Gass, Susan M. and Neu, Joyce (eds.): *Speech Acts Across Cultures: Challenges to Communication in a Second Language.* Berlin/ New York: Mouton de Gruyter, 65-86.

Blum-Kulka, Shoshana; House, Juliane & Kasper, Gabriele (1989a): "Appendix: The CCSARP Coding Manual". In: Blum-Kulka, Shoshana; House, Juliane & Kasper, Gabriele (eds.): *Cross-Cultural Pragmatics: Requests and Apologies*. Ablex: Norwood, 273-294.

Blum-Kulka, Shoshana; House, Juliane & Kasper, Gabriele (1989b): "Investigating Cross-Cultural Pragmatics: An Introductory Overview". In: Blum-Kulka, Shoshana; House, Juliane & Kasper, Gabriele (eds.): *Cross-Cultural Pragmatics: Requests and Apologies*. Ablex: Norwood, 1-34.

Breuer, Anja & Geluykens, Ronald (2007): "Variation in British and American English requests. A contrastive analysis". In: Kraft, Bettina & Geluykens, Ronald (eds.): *Cross-Cultural Pragmatics and Interlanguage English*. München: Lincom Europa, 107-126.

Brown, Penelope & Levinson, Stephen (1987): *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Carter, Ronald & McCarthy, Michael (2006): *Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English. Grammar and Usage.* Cambridge: Cambridge University Press.

Chafe, Wallace L.; Du Bois, John W. & Thompson, Sandra (1991): "Towards a new corpus of spoken American English." In: Aijmer, Karin & Altenberg, Bengt (eds.): *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London/ New York: Longman, 64-82.

Clark, Herbert H. & Lucy, Peter (1975): "Understanding what is meant from what is said: A study in conversationally conveyed requests". In: *Journal of Verbal Learning and Verbal Behavior* 14 (1), 56-72.

Clark, Herbert H. & Brennan, Susan E. (1991): "Grounding in communication". In: Resnick, Lauren B.; Levine, John M. & Teasley, Stephanie D. (eds.): *Perspectives on Socially Shared Cognition*. Washington: APA Books.

Edmondson, Willis & House, Juliane (1981): *Let's Talk and Talk about it. A Pedagogic Interactional Grammar of English*. München etc.: Urban and Schwarzenberg.

Fraser, Bruce (1974): "An analysis of vernacular performative verbs". In: Shuy, Roger W. & Bailey, Charles-James (eds.): *Towards Tomorrow's Linguistics*. Washington: Georgetown University Press, 139-158.

Golato, Andrea (2003): "Studying compliment responses. A comparison of DCTs and recordings of naturally occurring talk". In: *Applied Linguistics* 24 (1), 90-121.

Hancher, Michael (1979): "The classification of cooperative illocutionary acts". In: *Language and Society* 8, 1-14.

Hindelang, Götz (1978): *Auffordern. Die Untertypen des Aufforderns und ihre sprachlichen Realisierungsformen*. Göppingen: Verlag Aufbau Kümmerle.

Holtgraves, Thomas (2007): "Automatic intention recognition in conversation processing". In: *Journal of Memory and Language* 58, 627-654.

Jautz, Sabine (2008): "Gratitude in British and New Zealand radio programmes. Nothing but gushing?". In: Schneider, Klaus P. & Barron, Anne (eds.): *Variational Pragmatics. A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam/ Philadelphia: Benjamins, 141-178.

Jucker, Andreas (2009): "Speech act research between armchair, field and laboratory: The case of compliments." In: *Journal of Pragmatics* 41 (8), 1611-1635.
Jucker, Andreas; Schneider, Gerold; Taavitsainen, Irma & Breustedt, Barb (2008): "Fishing for compliments. Precision and recall in corpus-linguistic compliment research". In: Jucker, Andreas & Taavitsainen, Irma (eds.): *Speech Acts in the History of English*. Amsterdam/ Philadelphia: Benjamins, 273-294.
Kohnen, Thomas (2008): "Tracing directives through text and time. Towards a methodology of a corpus-based diachronic speech-act analysis". In: Jucker, Andreas & Taavitsainen, Irma (eds.): *Speech Acts in the History of English*. Amsterdam/Philadelphia: Benjamins, 295-310.
Koike, Dale (1996): "Transfer of pragmatic competence and suggestions in Spanish foreign language learning". In: Gass, Susan & Neu, Joyce (eds.) (1996): *Speech Acts Across Cultures. Challenges to Communication in a Second Language.* Berlin: Mouton de Gruyter: 257-81.
Koike, Dale (1994): "Negations in Spanish and English suggestions and requests: Mitigating effects?" In: *Journal of Pragmatics* 21, 513-526.
Leech, Geoffrey & Svartvik, Jan (1994): *A Communicative Grammar of English*. 2nd edition. London/ New York: Longman.
Manes, Joan & Wolfson, Nessa (1981): "The compliment formula". In: Coulmas, Florian (ed.): *Conversational Routine. Explorations in Standardized Communication Situations and Prepatterened Speech*. The Hague, etc.: Mouton, 115-132.
Martínez Flor, Alicia (2004): The effect of instruction on the development of pragmatic competence in the English as a foreign language context: A study based on suggestions. Unpublished doctoral dissertation. University of Jaume I (Department of English Studies).
McCarthy, Michael (2002): "Good listenership made plain. British and American non-minimal response tokens in everyday conversation". In: Reppen, Randi; Fitzmaurice, Susan M. & Biber, Douglas (eds.): *Using Corpora to Explore Linguistic Variation*. Amsterdam/Philadelphia: Benjamins, 49-71.
McCarthy, Michael & Carter, Ronald (2004): "Introduction". In: *Journal of Pragmatics* 36, 147-148.
O'Keeffe, Anne & Adolphs, Svenja (2008): "Response tokens in British and Irish discourse. Corpus, context and variational pragmatics". In: Schneider, Klaus P. and Barron, Anne (eds.): *Variational Pragmatics. A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam/ Philadelphia: Benjamins, 69-98.
Rintell, Ellen (1979): "Getting your speech act together: The pragmatic ability of second language learners". *Working Papers on Bilingualism* 17, 97-106.
Sadock, Jerrold (1974): *Toward a Linguistic Theory of Speech Acts*. New York, etc.: Academic Press.
Searle, John (1975): "Indirect speech acts". In: Cole, Peter & Morgan, Jerry (eds.): *Syntax and Semantics*. *Vol. 3: Speech Acts*. New York etc.: Academic Press, 59- 82.
Searle, John (1976): "A classification of illocutionary acts". In: *Language and Society* 5, 1-23.
Searle, John (1979): *Expression and Meaning*. Cambridge: Cambridge University Press.
Trosborg, Anna (1995): *Interlanguage Pragmatics. Requests, Complaints and Apologies*. Berlin/ New York: Mouton de Gruyter.
Tsui, Amy (1994): *English Conversation*. Oxford: Oxford University Press.
Yuan, Yi (2001): "An inquiry into empirical pragmatics data-gathering methods: Written DCTs, oral DCTs, field notes, and natural conversations". In: *Journal of Pragmatics* 33 (2), 271-292.

Anaphoric Relations in the Copenhagen Dependency Treebanks

Iørn Korzen and Matthias Buch-Kromann
Copenhagen Business School

Abstract

The Copenhagen Dependency Treebanks (CDT) are a set of parallel treebanks for Danish, English, German, Italian, and Spanish. One of the main objectives of the CDT is to arrive at a unified description and annotation system for syntax, morphology, discourse, and anaphora. The treebanks are currently in the process of being annotated for these levels in all five languages. After a brief discussion of the subdivisions of the so-called bridging anaphors proposed by different scholars, we describe the classification and terminology adopted in the CDT. The main distinction here is the very common one between coreferential and associative anaphors, special attention being given to the latter group. Resumptive and evolving anaphors are treated as special subgroups of the coreferential anaphors. A list of the associative relations proposed by the CDT with authentic examples concludes the paper.

1 Introduction. The Copenhagen Dependency Treebanks

The purpose of this paper is partly to discuss the classification system and terminology adopted for anaphora by various scholars, and partly to describe the way anaphora is treated in the Copenhagen Dependency Treebanks. Special attention will be given to the "associative anaphors", which appear to be the most complex of the main anaphor types.

The Copenhagen Dependency Treebanks, CDT, are a set of parallel treebanks for Danish, English, German, Italian, and Spanish which are currently being annotated for syntax, morphology, discourse, and anaphora in all five languages.[1] The corpus consists of 100,000 words compiled from 200-250 word excerpts from Danish mixed-genre texts, which have been translated into the other languages by native translators. All 100,000 words have been translated into English, while 70,000 words have been translated into each of the other languages. All texts have been automatically annotated for parts of speech. A main objective of the CDT is to arrive at a unified description and annotation system for syntax, morphology, and discourse which at the same time can take cross-linguistic differences into account (Buch-Kromann et al. 2009).

After a brief terminological discussion in section 2, section 3 describes the distinction between the main anaphor types adopted in the CDT, and section 4 presents the CDT analysis of coreference. Sections 5 and 6 are dedicated to associative anaphora and section 7 to a few technical remarks.

2 "Bridging", "coreferential" and "associative" anaphors

The terms "bridge" and "bridging" (in the sense relevant to this paper) probably first appear in Clark (1975). Here, bridging is defined as the construction of the implicatures with which the listener bridges "the gap from what he knows to the intended Antecedent" (Clark, 1975, 170). Clark includes "direct reference" (possibly

1 At this point, the anaphora analysis and annotation are confined to nominal anaphora.

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 83-98, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

with same-head NPs), "indirect reference by association", "indirect reference by characterization" (i.e. semantic roles), and the rhetorical relations "reasons", "causes", "consequences", and "concurrences" as situations that require an implicature "of some sort".

Subsequently, the term "bridging" has appeared frequently in the linguistic and computational literature, with more or less the same subclasses, except that coreferential pronouns and same-head NPs are generally left out, see e.g. Poesio *et al.* (1997, 2), Vieira and Poesio (2000, 558), and Caselli (2009, 73). In Vieira and Poesio (2000, 542), the "bridging descriptions" are summed up to be the "definite descriptions that either

(i) have an antecedent denoting the same discourse entity, but using a different head noun (as in *house . . . building),* or

(ii) are related by a relation other than identity to an entity already introduced in the discourse".

The same distinction, but expressed with the terms "coreferential" and "associative anaphors" respectively, is found in the work of a number of scholars, especially in the Romance tradition. Poesio and Vieira (1998, 187) cite Hawkins' (1978, 107/123) distinction between "Anaphoric Uses" and "Associative Anaphoric Uses", but in French the term "associative" is actually a lot older. It was probably first used as early as 1919 by Guillaume (1919, 162-163) but is now generally found in the theoretical linguistic literature, see e.g. Kleiber (1997a/b; 2001), Schnedecker et al. (1994), Cornish (1999), Lundquist (2000), Korzen (2003; 2009), and many others.[2]

In the last decade, also a number of schemes for anaphoric annotation have been released. Some of them confine themselves to coreference relations, e.g. the VENEX corpus (Poesio et al., 2004), the Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos, 2007), and the Portuguese and French corpus analysed by Vieira et al. (2002). On the other hand, the analyses e.g. of the GNOME Corpus (Poesio, 2004), the ARRAU Corpus (Poesio and Artstein, 2008), the Dutch COREA corpus (Hendrickx et al., 2008), and the Italian Live Memories Corpus (Rodríguez et al., 2010) consider coreference as well as certain associative relations such as set membership, subset, ownership, and part-of relations. The Prague Dependency Treebank, PDT (Nedoluzhko et al. 2009) performs a wider range of bridging annotation including relations such as contrast, location–resident, relatives, and event–argument. Navarretta (2010) focuses on abstract pronominal anaphora in the DAD parallel corpora. Unlike most of the cited studies, which use automatic or semi-automatic annotation for instance of "markables", i.e. text constituents (mainly NPs and pronouns or pronominal phrases) that may enter in anaphoric relations,[3] all anaphor annotation is done manually in the CDT.

3 Main anaphor types in the CDT

In view of the very obvious differences between coreferential and associative relations, the same overall distinction has been retained in the CDT, whose aim it is

2 For other designations of the "associative" anaphors in the theoretical linguistic literature, see e.g. Korzen (1996, 548-549).

3 The PDT has explicitly chosen not to use markables.

to handle all nominal anaphor types in the two groups, thus, among the coreferential types, both same-head and non-same-head NP anaphors.

Graphically, the relation between text constituents and discourse referents (DRs)[4] in the two cases may be described as in Figure 1, where the dashed arrows indicate the "bridging", or relation deduction, undertaken by the hearer/reader, and the dotted double arrow in the case of the associative anaphor (part B of the Figure) indicates the "association" between the two discourse referents in question:

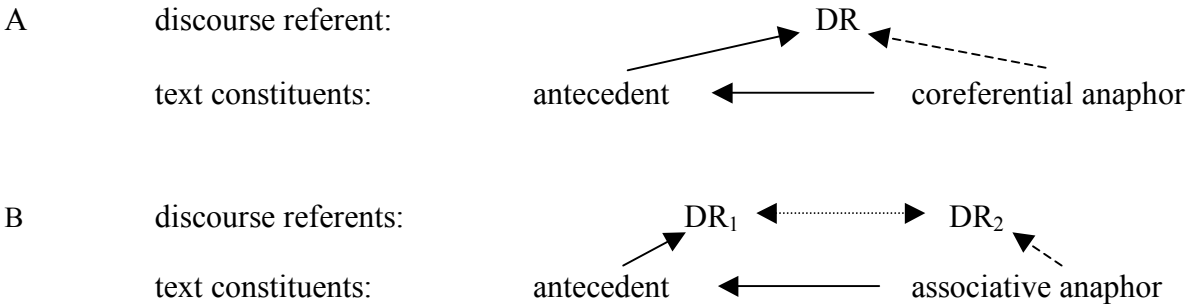


Figure 1. The relation between text constituents and discourse referents in the case of coreferential and associative anaphors.[5]

The so-called "evolving" anaphors refer to the same discourse referent as the antecedent, but after it has undergone radical changes in its ontological status, e.g.:

(1) The compactor crushed *a VW*. A huge crane then moved ***it*** to a railcar. (cit.: Asher, 2000, 142).[6]

Therefore they can be seen as a sort of interface between coreference and associative anaphors, since the discourse referent is technically the same but markedly different.

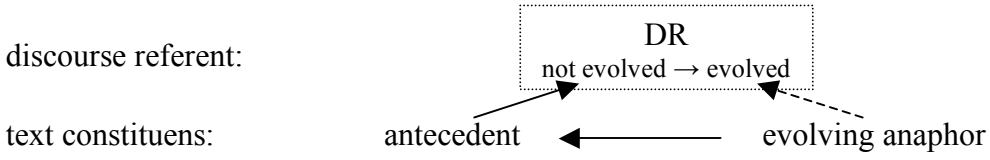


Figure 2. The relation text – extralinguistic context in the case of evolving anaphors.

In order to restrict the number of main anaphor types, the CDT treat evolving anaphors, as well as resumptive anaphors (which anaphorise whole sentences, clause or predicates, see Table 1 below and footnote 10), as special coreferential *sub*groups.

4 Coreference in the CDT

In the CDT annotation, coreferential anaphors are subdivided partly according to their linguistic material (pronouns, same-head NPs and non-same-head NPs) and partly according to their semantic content (resumptive and evolving anaphors). The

4 Discourse referents in the sense proposed by Karttunen (1969) and since then widely adopted in the literature.

5 In Webber's (1988) terminology, the coreferential anaphor specifies the referent which has already been evoked and specified by the antecedent, whereas the associative anaphor specifies **and** evokes its referent.

6 On "evolving anaphors" see also for instance Charolles and Schnedecker (1993), Korzen (2006), and Lundquist (2007).

CDT annotation arrows go from antecedent to anaphor and – in the case of a longer anaphoric chain – from the last occurring anaphor to the new one:[7]

Coreferential anaphor (and cataphor) labels	**Examples (antecedent → anaphor)**
A. COREF: coreferential pronouns and other pro-forms	*a car → it/this; John Smith → he; you → you;*[8]
B. COREF-IDEN: coreferential NPs with lexical identity	*a car → the car; a big car → the/this big car;*
C. COREF-VAR: coreferential NPs with lexical variety	*a car → the vehicle; a yellow car → the/this car;*[9]
D. COREF-RES: resumptive anaphors	[a sentence, clause or predicate] → *the episode, this incident*[10]
E. COREF-RES.PRG: resumptive anaphors referring to a speech act	*"I shall be back tomorrow" → the threat, the promise, the statement*
F. COREF-EVOL: evolving anaphors	see example (1) above

Table 1. Coreferential anaphor (and cataphor) types.

See also Figure 5 in section 7. The distinction between A, B and C concerns the linguistic material, whereas D, E and F are special semantic subgroups. D and E will also be either COREF (in the case of a pronoun) or COREF-VAR (in the case of an NP), and F will be either COREF, COREF-IDEN or COREF-VAR,[11] but this is not specified in the annotation. Cases of repeated proper nouns, e.g.:

(2) a. *John Smith → John Smith*
b. *John Smith → John*

are included as cases of COREF-IDEN (2a) or COREF-VAR (2b), even if they differ from common noun anaphors by not being necessarily dependent on their antecedent. Such cases, as well as repeated deictic pronouns, can easily be found and studied in searches that combine the anaphor label and the part-of-speech label.[12]

Our COREF-VAR group is very heterogeneous, at this point in time containing both cases of different (common or proper) nouns and different attributives. In due course, we may decide to subdivide this group into more homogeneous subgroups.

5 Association and the Generative Lexicon

As is well-known, the ways in which two discourse referents may "associate", as illustrated by the dotted double arrow in Figure 1B above, have been discussed

7 The subgroups of Table 1 include both anaphors and cataphors. In the case of cataphors, the arrows go from postcedent to cataphor.

8 Similarly, we annotate anaphoric relations between the subject of a verb of saying and a coreferential pronoun in the direct or indirect speech, and between coreferential pronouns in the different parts of dialogue, e.g. A: *...I...* → B: *...you...*

9 If, in a longer anaphoric chain, there is a relation between a pronoun, e.g. *he* (as last occurring anaphor), to an NP, e.g. *John Smith* (as new anaphor), this relation will be labelled COREF-VAR.

10 Typical resumptive anaphors are e.g. nominalizations, gerunds, and scene descriptions, and they can be subdivided in neutral NPs (e.g. *the operation, the activity, the situation*), NPs that either interpret or evaluate the story line (e.g. *the damage, the misdeed, the error*), or NPs that refer to the plot or story structure (e.g. *the scene, the gag, the comedy*); for more detail see also Korzen (2007).

11 On the material of the evolving anaphors, see especially the references mentioned in footnote 6. For more detail on the CDT annotation system, including the CDT manual, see the references in footnote 22.

12 Regarding the CDT search possibilities with the aid of the DTAG annotation tool, see Buch-Kromann *et al.* (2009).

extensively in the literature in the last few decades (as well as by Guillaume, 1919, 162ff.). Especially after the appearance of Pustejovsky's (1995) "Generative Lexicon", a number of scholars have seen the prospects of uniting lexical generativity, or "entailments" (Bos et al., 1995, 2), with the phenomenon of associative anaphors; see e.g. Bos et al. (1995), Lundquist (2000); Henry and Bassac (2008); Caselli (2009); Korzen (2000; 2003; 2009). Particularly useful is Pustejovsky's "qualia structure" with the four qualia, or roles, attributable to any artefact[13]:

A. FORMAL: That which distinguishes the object within a larger domain (orientation, magnitude, shape, dimensionality, color, position).
B. CONSTITUTIVE: The relation between an object and its constituents, or proper parts (material, weight, parts and component elements).
C. AGENTIVE: Factors involved in the origin or "bringing about" of the object.
D. TELIC: Purpose and function of the object.

Figure 3. Pustejovsky's (1995, 76ff. / 85ff.) "Qualia Structure".

Each of the four roles contains either entities/elements (A-B) or events (C-D) potentially generateable in a "default form", and both such entities/elements and events on the one hand and the arguments of the events on the other may be activated in an association relation to the object in question, i.e. function as associative anaphors to the NP designating this object.

But also before Pustejovsky, there were similar attempts to combine an apparent lexical and cognitive associability between concepts. For instance Hawkins (1978, 123-124) mentions "part-of relationship" and "attributes of an object" as possible "triggers" of associative anaphoric relations, such as for instance – with reference to *a car – the wheels, the steering-wheel, the passenger seats* and *the length, the colour, the weight,* corresponding to Pustejovsky's constitutive and formal qualia respectively.

Löbner (1998) distinguishes between sortal, relational and functional concepts, the functional ones being those that denote a 1-to-1 relation to a referent. He adds (Löbner, 1998, 4) that "all sortal nouns also encode relational or functional characteristics". Even a prototypical sortal noun like *book* has "a meaning that relates its possible referents to ways in which one can interact with books: write them, read them, […] etc.", a meaning that corresponds to Pustejovsky's agentive and telic qualia. Functional Concepts which have a possessor argument[14] are claimed to underlie all definite associative anaphors, whereas relational nouns which do not denote a 1-to-1 relation, e.g. *finger, hand, son, aunt*, are rejected as possible associative anaphors (Löbner, 1998, 10-11). This, however, is not necessarily true, as the following (very typical) Italian and Danish examples will show:

(3) Disse tutto questo e altro, che non ricordo. Mentre parlava, neppure io lo guardavo. [...] D'un tratto *mi* posò ***la mano*** su***l braccio***. "Avrei bisogno che tu mi dessi un consiglio",

13 Natural objects have a FORMAL and a CONSTITUTIVE quale, but not an AGENTIVE or a TELIC quale.
14 Such concepts are said always to have a situational argument as well and are therefore termed FC2s (Löbner 1998, 5).

fece. (Giorgio Bassani, *Gli occhiali d'oro.* Oscar Mondadori, Verona, 1973, p. 139)
'He said this and other things that I don't recall. As he spoke, I didn't even look at him. […] [lit.:] Suddenly *me* he put ***the hand*** on ***the arm***. "I need you to give me some advice", he said.'

(4) Politiet affyrede to skud mod *manden*. Det ene ramte *ham* i ***låret***. (Danish TV2-news 17.4.99)
'The police fired two shots at *the man*. One hit *him* in ***the thigh***.'

Even if *la mano* 'the hand'*, il braccio* 'the arm' and *låret* 'the thigh' are all in the singular, there is no reason to believe that the people involved are mutilated, and certainly an expression such as *hit him in* followed by a singular form of a noun denoting body parts we (normally) have more than one of is very common in English as well, as a few searches on Google reveal.

Associative anaphors tend to appear particularly often in the Romance languages, where for instance a case such as (5) is quite common:

(5) In questo momento *Fiorenza* non c'è. Io sono ***la figlia***.
[lit.:] 'At the moment *Fiorenza* is not here. I am ***the daughter***.'

Example (5) is based on an authentic example cited and discussed in Korzen (1996, 518, see also p. 35-36), and in "real life" the person "Fiorenza" actually has two daughters.

Kleiber (1997a/b; 2001, 263-367) operates with the following typology of four main groups of associative anaphors:

A. MERONYMIC: the anaphor is a fixed part of the antecedent, e.g. *a car → the wheel, a cup → the handle.*
B. LOCATIVE: the anaphor is located in the antecedent, e.g. *a village → the church, a kitchen → the refrigerator.*
C. FUNCTIONAL: the anaphor fulfils a function in relation to the antecedent, e.g. *a town → the mayor, a restaurant → the waiter.*
D. ACTANTS: the anaphor has a semantic and/or syntactic role in relation to a predicative antecedent, e.g. *an operation → the surgeon / the patient* (arguments)*, he cut the bread and put away* ***the knife*** (instrument).

Figure 4. Kleiber's (1997a/b; 2001, 263-367) typology of associative anaphors.

Of these, the A, C and some of the D types are covered by Pustejovsky's qualia. One could argue that some of Kleiber's types are overlapping: a meronymic anaphor is also located in the antecedent and may very well fulfil a function as well. We shall return to these (thorny) problems below.

6 The associative anaphors in the CDT

The CDT classification and subdivision of associative anaphors are highly inspired by the typologies mentioned in the previous section, as the following lists will show. Since CDT is an ongoing project in which we by and large have worked, and are working, empirically, we cannot exclude that further analyses will give rise to changes, but we believe they will be minor. In the following text examples (all authentic), the antecedents are printed in italics and the anaphors in bold italics followed by the label. A number between parentheses following the example indicates the number of the text in the CDT corpus. In a few cases, text examples come from other sources. Unlike the coreferential anaphors, the structure of the associative labels is hierarchic, which means that the following types are all associative subtypes. As in the CDT syntax and discourse annotation (and inspired by the Penn Discourse Treebank), this means that in case of uncertainty, the annotator can remain on a higher (more generic) annotation level.[15]

With a few exceptions (see footnote 16), associative anaphors seem classifiable according to two parameters:

- lexical semantics and generativity, qualia structure;
- semantic roles in relation to a predicate; the predicate may be either directly expressed by the antecedent or generatable from it.

1. Qualia structure	2. Semantic roles	3. Other types[16]
ASSOC-FORMAL	ASSOC-AGENT	ASSOC-LOC(ation)
ASSOC-CONST(itutive)	ASSOC-PATIENT	ASSOC-TIME
ASSOC-AGENTIVE	ASSOC-EXPER(iencer)	ASSOC-EVENT
ASSOC-TELIC	ASSOC-REC(ipient)	
	ASSOC-INST(rument)	

Table 2. Associative subtypes, parameters and labels.

6.1 The anaphor is associated with the antecedent with regard to its qualia structure

ASSOC-FORMAL

The FORMAL quale expresses static information about the object's characteristics. If the anaphor is associated with the antecedent with regard to its FORMAL quale, it may designate the shape, dimension, colour, etc. of the object designated by the antecedent:

(6) The ham to be used in the dish must not be too salty. You cannot use *the thin slices*, which are packaged in the refrigerated counter. *They* are too salty and too wet and ***the flavour*** [ASSOC-FORMAL] is not good enough. (148)

The other three qualia roles contain information about relations that the object

15 A similar solution does not seem to be needed in the case of the coreferential anaphors, where our subdivision should not give rise to much uncertainty. At the most, a COREF-RES or a COREF-EVOL anaphor might risk a categorization as a COREF, a COREF-IDEN or a COREF-VAR anaphor, which would neither be catastrophic, nor untrue.

16 In fact, these "exceptions" may be seen as extensions of the other two subtypes. However, LOCATION and TIME are labelled as semantic roles by some scholars, see e.g. Larson (1984, 202), and EVENT expresses a predication linked to the antecedent, similar to but more generic than the TELIC and AGENTIVE qualia. See 6.3 below.

referred to by the antecedent can be a part of, i.e., they constitute predicates of which the antecedent is an argument. In these cases, an associative anaphor can function as the other argument or as the predicate itself.

ASSOC-CONST

Also the CONSTITUTIVE quale expresses static information about the object (parts, elements, material, content, etc.). The predicates of which antecedent and anaphor are arguments are *has_part, consists_of, is_part_of,* and the like. In (7) the anaphor is part of the antecedent, in (8) vice versa. In both cases we talk about ASSOC-CONST-relations:[17]

(7) The accident took place at dinner time around 6:45 p.m. last night […]. I saw *the plane* with its nose pointing downward, ***the left wing*** [ASSOC-CONST] up and ***the right wing*** [ASSOC-CONST] down over behind the flat building. (1536)

(8) On September 8, DE BEERS CENTENARY opened an office in *Moscow*. Present were also De Beers' top people, Russian politicians, diplomats and representatives of ***the country***'s [ASSOC-CONST] diamond industry and trade. (431)[18]

ASSOC-TELIC and ASSOC-AGENTIVE

If the anaphor is associated with the antecedent with regard to its AGENTIVE or TELIC quale, the anaphor may designate the quale predicate itself or an inferable argument of such a predicate. Examples (9) and (10) are cases of predicative anaphors:

(9) As previously explained, we were waiting for an approval from Sony as we submitted to them *a new version of Blood Bowl PSP*. […] *This new version* has been finally approved and ***the production*** [ASSOC-AGENTIVE] started. Please find below the list of fixes that were made. (http://www.gamefaqs.com/boards/944028-blood-bowl/52159350, accessed October 8th, 2010)

(10) However, not all debriefings are held after the simulation, but in *certain instances*, for example, where ***the aim*** [ASSOC-TELIC] is to teach a technical skill […] debriefing may occur during the simulation, in-scenario debriefing.
(citeseerx.ist.psu.edu/viewdoc/download, accessed October 13th, 2010)

Anaphors that designate a particular semantic role of the given quale predicate are treated as subtypes. The precise analysis of the role in question will depend on the inferred predicate. Thus, in these cases the annotators are asked to add the inferred predicate between parentheses. As regards AGENTIVE subtypes, so far we have only encountered the semantic role AGENT:

(11) In April 2003, marking the tenth anniversary of the Waco Massacre, *a new film* was released. According to ***the producer*** [ASSOC-AGENTIVE.AGENT/(produce)], "Waco: A New Revelation" is a film so disturbing that […] it triggered new investigations in both

17 "The constitutive […] quale refers not only to the parts or material of an object, but defines, for an object, what that object is logically part of, if such a relation exists. The relation PART-OF allows for both abstractions" (Pustejovsky 1995, 98).

18 It may be debatable whether the antecedent is *Russian* rather than *Moscow,* but they can both function as antecedents, which can be proved in a simple test where one or the other is omitted from the co-text. We should also add that in cases like this, a precise borderline between ASSOC-CONST and ASSOC-LOC can be very hard to draw; see below.

houses of Congress […]. (http://www.serendipity.li/waco.html, accessed September 5th, 2010)

In (11), in order to infer *the producer*, we must first activate the agentive quale *produce.* Similarly, in (12) and (13) *the pilot* and *both apprentices* can be seen as the semantic roles AGENT and PATIENT of the telic qualia of a *flight* (i.e. *to fly*) and a *test* (i.e. *to examine*) respectively:

(12) The accident took place at dinner time around 6:45 p.m. last night, shortly after *the El-Al flight* […] lifted off from Amsterdam's Schiphol airport.
The pilot [ASSOC-TELIC.AGENT/(fly)] suddenly reported to the control tower that he had engine problems […]. (1536)

(13) *Two journeyman tests* were passed in August. ***Both apprentices*** [ASSOC-TELIC.PATIENT/(examine)] are trained at the Royal Copenhagen A/S Georg Jensen Silversmithy. (431)

In some cases, more than one subtype interpretation may apply, for which reason an annotator could remain at the ASSOC-TELIC level[19]:

(14) The men in question are simply film reviewers and quite harmless. […] If some nonsense should sometimes appear in a *film review*, it is thus due not to time pressure, even though, of course, it is most convenient for the reviewers if ***the readers*** [ASSOC-TELIC.AGENT/(read) or ASSOC-TELIC.REC/(receive)] believe that. (647)

6.2 The antecedent is predicative and the anaphor is a semantic role

If the antecedent is a predicate or a predicative noun, the anaphor can constitute a semantic role which is related to it directly, not (necessarily) via a quale:

(15) *The operation* itself requires general anesthesia ... the patient is asleep for the entire course of the operation. ***The surgeon*** [ASSOC-AGENT] opens the chest by dividing the breast bone or sternum. (http://www.heartsurgeons.com/pr3.html, accessed August 5th, 2010)[20]

(16) *The operation* itself requires general anesthesia ... ***the patient*** [ASSOC-PATIENT] is asleep for the entire course of the operation. The surgeon opens the chest by dividing the breast bone or sternum. (http://www.heartsurgeons.com/pr3.html, accessed August 5th, 2010)

(17) *The accident* took place at dinner time around 6:45 p.m. last night […]. "[…] The pilot attempted to right the plane - then I could not see more, but suddenly there were sparks in the air," says ***eyewitness Peter de Neef*** [ASSOC-EXPER]. (1536)

(18) "[…] This is *the most violent attack* to this point. ***The bombs*** [ASSOC-INST] fell half a mile from the hotel," reported John Hollimann […] (61).

19 With the risk of confusion with cases such as (10). At this point, we have not been able to solve this problem.
20 The tree dots appeared as shown in the cited text.

6.3 Other types

According to the definition of "semantic roles" (see footnote 16), TIME and LOCATION may belong to the previous section or they may be extensions of it. An ASSOC-TIME anaphor may indicate a point in time linked to the antecedent, which may be a predicate or predicative noun, another time indication, as in (19), or a more general narrative frame, as in (20):

(19) As mentioned, the season will begin on *March 16* with the showdown between AGF and Brøndby, followed ***the day after*** [ASSOC-TIME] by games between: Ikast-Lyngby, B 1903-Silkeborg, AaB-Vejle and FremOB. (43)

(20) Aspiring chef dies hours after making ultra-hot sauce for chilli-eating contest [headline]
Andrew Lee made an ultra-hot sauce with homegrown chillis. ***The morning after*** [ASSOC-TIME] he was found unconscious and paramedics were unable to revive him.
(Mailonline, http://www.dailymail.co.uk/news/, accessed August 6th, 2010)

The ASSOC-LOC relation is very close to the ASSOC-CONST relation, and a precise borderline can be hard to draw. An ASSOC-LOC anaphor is located in the antecedent (or vice versa) without being necessarily a constitutive part:

(21) Upon entry, the officers saw *the kitchen* with many dirty dishes, spoiled food on the floor and in ***the refrigerator***, and bags of trash and other combustibles on top of ***the stove***.
(http://www.leagle.com/xmlResult.aspx?xmldoc=197621858CalApp3d160_1205.xml&docbase=CSLWAR1-1950-1985, accessed November 10th 2010).

Similarly, as an extension of examples (9) and (10), a predicative anaphor may express an event which is associable with the antecedent, but not necessarily with regard to its qualia structure. In such cases we adopt the more generic label ASSOC-EVENT:

(22) Hamid Jafar was very eager to show his appreciation of the agreement to his *Iraqi* partners. Shortly before ***the invasion*** [ASSOC-EVENT], he ordered an engraved, Swiss, gold pistol assessed at 7,000 pounds from […] the English Queen's jeweller in London. (939)

7 Graphs and inter-annotator agreement

The CDT graphs are generated with the DTAG annotation tool described in Kromann (2003)[21] and use directed edges with the relation labels shown at the arrow head. Figure 5 shows the syntax annotation (above the nodes) and anaphor annotation (below the nodes) of the last sentence of example (7).

21 More references can be found at http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT.

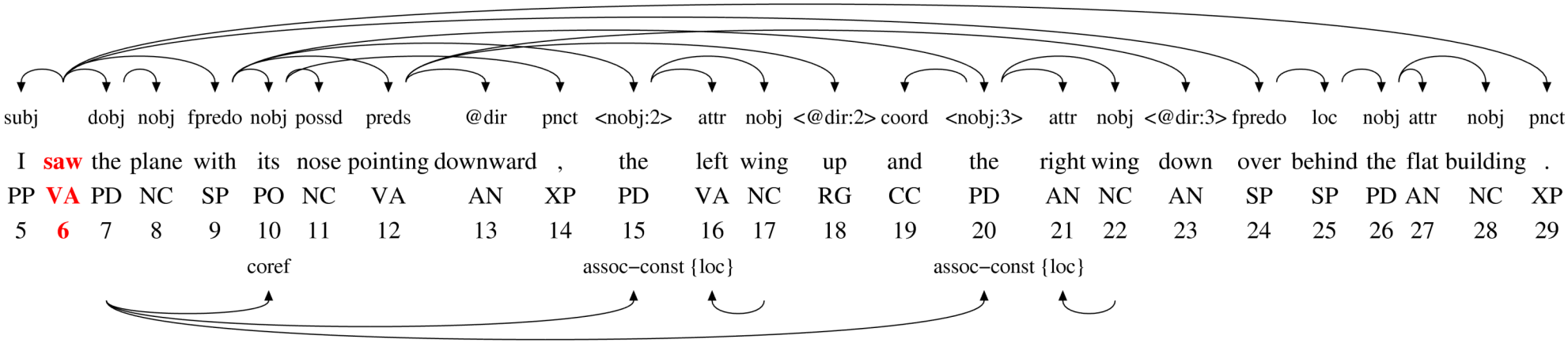


Figure 5. A CDT syntax and anaphor annotation of the sentence *I saw the plane with its nose pointing downward, the left wing up and the right wing down over behind the flat building.*

The annotation shows that the NP *the plane* (nodes 7-8)[22] is the antecedent of a coreferential pronoun (node 10) and two ASSOC-CONST anaphors (nodes 15-17 and 20-22). In this figure, we have omitted most of the secondary semantic relations, also annotated below the text, but left two of them behind (nodes 16 and 21) in order to give an impression of this annotation category. Even if they are annotated below the text like the anaphoric relations, secondary semantic relations clearly belong to a different linguistic dimension, just like the syntactic and discourse relations belong to different dimensions although they are both annotated above the text in the CDT.

In order to test our anaphor relation system by computing inter-annotator agreement as soon as possible, 25 texts have been annotated independently by two annotators. The texts contained a total of 466 anaphoric relations, and Table 3 (taken from the CDT manual, Buch-Kromann et al., 2010) shows the level of inter-annotator agreement and the frequency of the anaphoric relations found in the 25 texts. Agreement is reported as percentage agreement[23] in the following way:

- *Full labelled agreement, A:* the probability that another annotator assigns the same label and out-node to the relation;
- *Unlabelled agreement, A_U:* the probability that another annotator assigns the same out-node (but not necessarily the same label) to the relation;
- *Label agreement, A_L:* the probability that another annotator assigns the same label (but not necessarily the same out-node) to the relation.

Relation name	Agreement % $A - A_U - A_L$	Relation count	Relation name	Agreement % $A - A_U - A_L$	Relation count
COREF	84 – 85 –92	141	ASSOC (subtype)	39 – 83 – 39	9
COREF-VAR	71 – 79 –79	97	ASSOC-LOC	100 – 100 – 100	5
REF[24]	100 – 100 – 100	63	ASSOC-AGENTIVE	25 – 50 – 50	4
COREF-IDEN	77 – 83 – 81	53	ASSOC-EVENT	100 – 100 – 100	3
ASSOC-CONST	59 – 77 – 67	39	ASSOC-FORMAL	100 – 100 – 100	1
COREF-RES	65 – 73 – 72	25	COREF-RES.PRG	0 – 0 – 0	1
ASSOC-TELIC	71 – 88 – 83	24	COREF-EVOL	0 – 100 – 0	1
			TOTAL	77 – 84 – 84	466

Table 3. Inter-annotator agreement based on 25 CDT texts with 466 anaphoric relations.

As a first test at a relatively early point in time, and considering that we include all nominal anaphors, even the most complex associative types, we find the result acceptable. However, we feel confident that an even better result can be obtained after more time for discussion and analysis together with the two annotators.

22 Of which the determiner is considered head and the lexical noun nominal object, "nobj". For more detail on CDT graphs, analyses, and annotation, see Buch-Kromann *et al.* (2009, 2010). The CDT-manual can be downloaded from the URL of the latter reference: http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf.

23 The estimated level of agreement is defined as the probability that another annotator assigns the same label and/or out-node to the relation (this number may be inaccurate if the relation count is small). We do not report chance-corrected scores because they are harder to interpret and their usefulness is contested (Reidsma and Carletta, 2008; Buch-Kromann, 2010). For more detail, we refer our readers to the CDT manual.

24 REF regards syntactically determined coreference, typically used in relative clauses with a relative pronoun.

8 Conclusion

In this paper, we have described the anaphor annotation system in the Copenhagen Dependency Treebanks, an on-going project which is still in its relatively early stages. The over-all distinction is the very common one between coreference and associative anaphors, of which the latter group is clearly the most complex and complicated one. Associative anaphora has to do with how concepts relate to or associate with each other, and in this connection we have found it fruitful to look at lexical generativity and semantic association. A combination of Pustejovsky's qualia structure and the most common semantic roles (in Table 4 "semroles") played by arguments in connection with their predicates seems to be able to account for almost all cases of associative anaphora. The CDT project operates with a hierarchic label system that allows annotators to remain at a higher level in case of uncertainty as to subtypes. The ASSOC types and subtypes are the following:

ASSOC-QUALIA (± semrole subtype):	**ASSOC-SEMROLE:**	**ASSOC-OTHER:**
ASSOC-FORMAL	ASSOC-AGENT	ASSOC-EVENT
ASSOC-CONST	ASSOC-EXPER	ASSOC-LOC
ASSOC-AGENTIVE	ASSOC-INST	ASSOC-TIME
ASSOC-AGENTIVE.AGENT	ASSOC-PATIENT	
ASSOC-TELIC	ASSOC-REC	
ASSOC-TELIC.AGENT		
ASSOC-TELIC.EXPER		
ASSOC-TELIC.INST		
ASSOC-TELIC.PATIENT		
ASSOC-TELIC.REC		

Table 4. Associative anaphora in the CDT.

At a later stage, cross-linguistic alignment will allow us to compare anaphoric relations in our five languages with great accuracy. For instance, it will enable us to identify and precisely describe the considerable typological differences between associative relations in Romance and Germanic languages, some of which were briefly illustrated in examples (3) and (5).

9 Acknowledgments

This work was supported by grants from the Danish Research Council for the Humanities and the Copenhagen Business School. We thank Lotte Jelsbech Knudsen and Morten Gylling-Jørgensen for many fruitful discussions and the anonymous reviewers for their useful comments.

References

Nicholas Asher. Events, Facts, Propositions, and Evolutive Anaphora. In James Higginbotham, Fabio Pianesi and Achille C. Varzi (eds.). *Speaking of Events.* Oxford University Press, New York & Oxford, pages 123-150, 2000.

Johan Bos, Paul Buitelaar, and Anne-Marie Mineur. Bridging as Coercive Accommodation. In *Workshop on Computational Logic for Natural Language Processing (CLNLP)*, Edinburgh, 1995.

Matthias Buch-Kromann. Open challenges in treebanking: some thoughts based on the Copenhagen Dependency Treebanks. Invited paper at the Annotation and Exploitation of Parallel Corpora Workshop, Tartu, December 1-2, 2010.

Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. Uncovering the 'lost' structure of translations with parallel treebanks. In Inger M. Mees, Fabio Alves, and Susanne Göpferich, (eds.), *Methodology, Technology and Innovation in Translation Process Research. Copenhagen Studies in Language* 38, Samfundslitteratur, Copenhagen, pages 199-224, 2009.

Matthias Buch-Kromann, Morten Gylling-Jørgensen, Lotte Jelsbech Knudsen, Iørn Korzen, and Henrik Høeg Müller. *The inventory of linguistic relations used in the Copenhagen Dependency Treebanks. Technical report. (The CDT manual)*. Center for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, 2010. http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf

Tommaso Caselli. Using a Generative Lexicon Resource to Compute Bridging Anaphora in Italian. *Procesamiento del Lenguaje Natural* 42, pages 71-78, 2009.

Michel Charolles and Catherine Schnedecker. Coréférence et identité. Le problème des référent évolutifs. In *Langages* 112, pages 106-126, 1993.

Francis Cornish. *Anaphora, Discourse, and Understanding. Evidence from English and French.* Clarendon Press, Oxford, 1999.

Herbert H. Clark. Bridging. In R. C. Schank & B. L. Nash-Webber (eds.), *Theoretical Issues in National Language Processing*. MIT, 1975.

Gustave Guillaume. *Le problème de l'Article e sa solution dans la Langue française*. Librairie Hachette, Paris, 1919. [Réédition Libraire A.-G. Nizet, Paris / Les Presses de l'Université Laval, Quebec, 1975].

John A. Hawkins. *Definiteness and Indefiniteness. A Study in Reference and Grammaticality Prediction.* Croom Helm, London, 1978.

Iris Hendrickx *et al.* A Coreference Corpus and Resolution System for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 144-149, 2008.

Patrick Henry and Christian Bassac. A toolkit for a Generative Lexicon. In *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris 2007, 2008.

Lauri Karttunen. Discourse Referents. In *International Conference on Computational Linguistics, COLING, Preprint No. 70*, 1969.

Georges Kleiber. Des anaphores associatives méronymiques aux anaphores associatives locatives. In *Verbum* XIX/1-2, pages 25-66, 1997a.

Georges Kleiber. Les anaphores associatives actantielles. In *Scolia* 10, pages 89-120, 1997b.

Georges Kleiber. *L'anaphore associative*. Presses Universitaires de France, Paris, 2001.

Iørn Korzen. *L'articolo italiano fra concetto ed entità. Vol. I-II*. [*Etudes Romanes* 36], Museum Tusculanum Press, Copenhagen, 1996.

Iørn Korzen. Pragmatica testuale e sintassi nominale. Gerarchie pragmatiche, determinazione nominale e relazioni anaforiche. In Korzen and Marello (eds.), 2000, pages 81-109, 2000.

Iørn Korzen. Anafora associativa: aspetti lessicali, testuali e contestuali. In Nicoletta Maraschio and Teresa Poggi Salani (eds.). *Italia linguistica anno Mille, Italia linguistica anno Duemila.* Bulzoni, Roma, pages 593-607, 2003.

Iørn Korzen. Tipologia anaforica: il caso della cosiddetta "anafora evolutiva". In *Studi di grammatica italiana.* Accademia della Crusca, Firenze, XXV, pages 323-357, 2006.

Iørn Korzen. Linguistic typology, text structure and anaphors. In Korzen and Lundquist (eds.),

2007, 93-109.
Iørn Korzen. Anafora associativa: ulteriori associazioni. In Federica Venier (ed.). *Tra pragmatica e linguistica testuale. Ricordando Maria-Elisabeth Conte.* [*Gli argomenti umani* 13]. Edizioni dell'Orso, Alessandria, pages 307-326, 2009.
Iørn Korzen and Carla Marello (eds.). *Argomenti per una linguistica della traduzione / On linguistic aspects of translation / Notes pour una linguistique de la traduction. Gli argomenti umani* 4. Edizioni dell'Orso, Alessandria, 2000.
Iørn Korzen and Lita Lundquist (eds.). *Comparing Anaphors. Between Sentences, Texts and Languages. Copenhagen Studies in Language,* 34. Samfundslitteratur Press, Copenhagen, 2007.
Olga Krasavina and Christian Chiarcos. PoCoS – Potsdam Coreference Scheme. In *LAW '07 Proceedings of the Linguistic Annotation Workshop*, 2007.
Matthias Trautner Kromann. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), 14-15 November, Växjö*, pages 217–220, 2003.
Mildred I. Larson. *Meaning-based translation. A guide to cross-language equivalence*. Lanham, New York / London, 1984.
Sebastian Löbner 1998. Definite Associative Anaphora. (manuscript) http://user.phil-fak.uni-duesseldorf.de/~loebner/publ/DAA-03.pdf
Lita Lundquist. Translating Associative Anaphors. A Linguistic and Psycholinguistic Study of Translation from Danish into French. In Korzen and Marello (eds.) 2000, 111-129, 2000.
Lita Lundquist. Comparing evolving anaphors in Danish and French. In Korzen and Lundquist (eds.), pages 111-125, 2007.
Costanza Navarretta. The DAD parallel corpora and their uses. In *Proceedings of LREC 2010, Malta, 17-23 May 2010*, pages 705-712, 2010.
Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*, pages 108–111, 2009.
Massimo Poesio. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*, 2004.
Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In Proceedings of the LREC 2008, Marrakech, Morocco, 2008.
Massimo Poesio and Renata Vieira. A corpus-based investigation of definite description use. In *Computational Linguistics* 24(2), pages 183-216, 1998.
Massimo Poesio, Renata Vieira and Simone Teufel. Resolving Bridging References in Unrestricted Text. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, Madrid, Spain, pages 1-6, 1997.
Massimo Poesio, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran, and Sara Tonelli. The VENEX corpus of anaphora and deixis in spoken and written Italian, 2004. http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf
James Pustejovsky. *The Generative Lexicon: A theory of computational lexical semantics.* MIT Press, Cambridge, MA, 1995.
Dennis Reidsma and Jean Carletta. Reliability measurement without limits. In *Computational Linguistics* 34(3), pages 319-326, 2008.
Kepa J. Rodríguez, Francesca Delogu,Yannick Versley, Egon W. Stemle, and Massimo Poesio. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, pages 157-163, 2010.
Catherine Schnedecker, Michel Charolles, Georges Kleiber, and Jean Davis (eds.). *L'anaphore associative. (Aspects linguistiques, psycholinguistiques et automatiques)*. Klincksieck, Paris, 1994.
Renata Vieira and Massimo Poesio. An Empirically-Based System for Processing Definite Descriptions. In *Computational Linguistics* 26(4), pages 539-593, 2000.

Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the DAARC, Estoril,* 2002.
Bonnie Webber. Tense as Discourse Anaphora. In *Computational Linguistics* 14 (2), pages 61-73, 1988.

Antecedent and Referent Types of Abstract Pronominal Anaphora

Costanza Navarretta
University of Copenhagen

Abstract

This paper is about the relation between pronominal types, syntactic types of the antecedent, semantic type of the referent and anaphoric distance in the Danish part of the DAD corpus comprising written and spoken data. These aspects are important to understand the use of abstract anaphora and to process them automatically and some of them have been investigated previously (see i.a. Webber (1988); Gundel et al. (2003); Navarretta (2010)). Differing from preceding studies, we extend the analysis of the syntactic types of the antecedent to include a fine-grained classification of clausal types and also investigate the anaphoric distance. The most common antecedent types in the data are *subordinate clause* and *simple main clause* and most abstract anaphora occurred in the clause which followed the antecedent or the clause in which the antecedent occurred. There is no clear dependence between the type of antecedent clause and the type of referent.

1 Introduction

In this paper we analyse the relation between abstract anaphora, syntactic types of antecedent, semantic types of referent and anaphoric distance in an annotated corpus of Danish texts, monologues and dialogues. Abstract anaphora indicate here third person singular pronouns which have as antecedents copula predicates, verbal phrases, clauses and discourse segments of varying size and refer to abstract types such as properties, events, situations and propositions. Abstract anaphors are also known as *impure textual deictics* (Lyons, 1977) and *discourse deictics* (Levinson, 1987; Webber, 1991) while reference to abstract entities independently of the type of antecedent has been called *situation reference* (Fraurud, 1992) and *abstract object reference* (Asher, 1993). An example of abstract anaphor in Danish is in (1) where the stressed pronoun *duet*[1] (this/that) has as antecedent the precedent utterance *She would certainly ensure that he came over there*:

(1) A: *Hun skulle nok sørge for han kom derover*
'She would certainly ensure that he came over there'

B: *d'et kunne jeg godt være sikker på*
'of THAT I could be completely sure'
(LANCHART)

Abstract anaphora are very frequent in languages such as English (Byron and Allen, 1998; Gundel et al., 2003) and Danish (Navarretta, 2000), but abstract pronominal reference varies from language to language (see i.a. Fraurud (1992); Borthen et al. (1997); Kaiser (2000); Navarretta (2002, 2010)).

Resolving abstract anaphora automatically is difficult because the antecedents belong to various syntactic types and have varying size. The anaphor can immediately

[1] We mark a stressed vowel with an apostrophe before its occurrence.

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 99-110, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

follow its antecedent, but it can also occur several clauses later. Furthermore there is not a one to one relation between the antecedent's syntactic type and the referent's semantic type (see Webber (1991); Gundel et al. (2003)).

Various algorithms for resolving abstract anaphora in English have been proposed (i.a. Eckert and Strube (2001); Byron (2002); Strube and Müller (2003); Müller (2007)). These algorithms rely on the pronominal type, on linguistic, semantic, domain-specific knowledge and/or on the annotations in domain-specific corpora. Their results are still not good enough to be used in practical applications and the evaluation of the algorithms indicates that the recognition of the anaphoric uses of the pronouns and the identification of the antecedents are some of the most problematic aspects. Although Danish anaphora have different characteristics than the English ones, the identification of the correct antecedent is also problematic in this language (Navarretta, 2002, 2004a).

The aim of the present work is to provide more knowledge about the use of abstract anaphora in Danish.

In previous investigations of abstract anaphora in Danish we have looked at the relation between type of pronouns, semantic referent types and syntactic types of antecedent. In the present work we extend the investigation to comprise a very fine-grained analysis of different types of clausal antecedents and to include anaphoric distance that is the number of clauses between the abstract anaphor and its antecedent. The anaphoric distance has been seen as one important salience indicator and it is the determining factor behind the accessibility hierarchy of nominal referring expressions proposed by Ariel (1988, 1994). Because the semantic type of the referent depends on the context in which the abstract anaphor occurs (see i.a. Webber (1991); Eckert and Strube (2001)), we believe that finding possible relations between these semantic types, the syntactic types of the antecedent, the anaphoric distance and the types of pronoun can contribute to the construction of resolution algorithms.

The paper is organised as follows. In Section 2 we discuss related work and in Section 3 we shortly describe the Danish abstract anaphora and the corpus upon which we base our research. In Section 4 we present and discuss the results of our investigation and in Section 5 we conclude.

2 Background work

Webber (1991) notices that abstract anaphors with the same antecedent can refer to objects of different semantic type depending on the context in which the anaphor occurs. She suggests that abstract pronouns create their referents in the moment they are uttered by an act of *ostension*.

The relation between the type of abstract pronoun and the syntactic type of the antecedent has been addressed by i.a. Webber (1988); Hegarty (2003); Gundel et al. (2003, 2004); Navarretta (2004, 2007, 2010). In particular Webber (1988) reports that

personal pronouns in English often cannot refer to abstract entities when the antecedent is a clause, because the clause is not accessible to the pronoun. In corpus-based studies of the occurrences of abstract anaphora in English Byron and Allen (1998), Gundel et al. (2003) and Hedberg et al. (2007) confirm Webber's observation .

Hegarty (2003) explains the frequency of occurrence of demonstrative pronouns with clausal antecedents in terms of the *Givenness Hierarchy* (Gundel et al., 1993). According to him entities introduced in discourse by clauses are only activated in the cognitive status of the addressee, while entities introduced in discourse by verbal phrases are similar to entities introduced in discourse by nominal phrases: they are often in focus and can be referred to by the personal pronoun *it*. Because the most common referent types when the antecedents are clauses are facts, situations and propositions, demonstrative pronouns refer much more often to facts, situation and propositions than personal pronouns do. On the other hand the referents with verbal phrases antecedents are events or states and they are thus often referred to by personal pronouns.

Navarretta (2007, 2010) reports that, differing from the English *it*, Danish and Italian personal pronouns have often clausal antecedents and thus in numerous cases they refer to facts, situations and propositions. She explains one of the differences in abstract reference between Danish, English and Italian in terms of the three languages' pronominal systems and syntactic structure. With respect to syntax she notices that in Danish constructions such as left dislocation and clefts are much more frequent than in English. These constructions put the clauses in focus in information structure terms because and thus the clauses are in these cases very salient, or *in focus* in terms of the *Givenness Hierarchy* (Gundel et al., 1993).

In our preceding studies we accounted for the characteristics of anaphora with nominal phrase and verbal phrase antecedents opposed to anaphora with clausal antecedents and on the differences between Danish, English and Italian. In the present work we focus on the Danish data and add to our investigation the analysis of the types of clausal antecedent and the anaphoric distance and their relation to the referent types and the pronominal types.

3 The data

3.1 Abstract anaphora in Danish

In Danish texts two abstract pronouns are used: the pronoun *det* (it/this/that) which is ambiguous with respect to its pronominal type and the demonstrative pronoun *dette* (this). In spoken Danish abstract pronouns comprise the unstressed personal pronoun *det* (it) and the stressed demonstrative pronouns *d'et* (this/that), *d'et h'er* (this) *d'et d'er* (that). The demonstrative pronoun *dette* (this) occurs extremely seldom in spoken language and in our data it only occurred two times and in both cases it had a nominal phrase antecedent.

3.2 The corpus

Our study is mainly based on the Danish part of the DAD corpora (Navarretta and Olsen, 2009) which consist of the following data:

- Transcriptions of the DANPASS corpus (Grønnum, 2006) which is the Danish version of the MAPTASK corpus (Anderson et al., 1991) and comprises both dialogues and monologues. The DANPASS dialogues contain 52,145 running words while the monologues consist of 21,224 words.
- Transcriptions of multiparty spontaneous dialogue extracts from the LANCHART corpus (Gregersen, 2007) comprising 24,112 running words.
- Transcriptions of two TV-interviews from the Danish public television *DR* (2,192 words).
- Translations from Italian of three Pirandello (1922-1937) stories (11,280 words).
- EU texts (24,389 words).
- Danish juridical texts (11,600 words).
- Extracts of newspaper and journal articles, novels and reports (12,570 words) from the Danish general language PAROLE corpus (Keson and Norling-Christensen, 1998).

3.3 The annotations

The DAD data contain many types of annotation such as structural information for the texts and speaker and turn information for the dialogues, PoS and lemma information, information about the functions and uses of third person singular neuter pronouns and especially their anaphoric uses (Navarretta, 2010). A description of the annotation schemes used and a report of inter-coder agreement measures for the various annotation types are in Navarretta and Olsen (2008); Navarretta (2009); Navarretta and Olsen (2009). The annotations which are relevant to the present work are the following:

- the type of pronoun, e.g. *det*, *dette*, unstressed *det*, stressed *det*;
- the antecedent;
- the syntactic type of the antecedent;
- the semantic type of the referent;
- the anaphoric distance in term of clauses.

The syntactic types of antecedent relevant to this work are *verbal phrase* (VP), *adjectival phrase*, *prepositional phrase* and *nominal phrase* in copula constructions (these three types are called *CPR* henceforth), *discourse segment* (DS) and *clause*. The

clausal type is furthermore distinguished in the following subtypes: *simple main clause* (CL) which covers main clauses which do not have subordinate clauses, *matrix clause* (MCL), *subordinate clause* (SCL) and *complex clause* (CCL) which comprise co-ordinated clauses and main clauses with their subordinated clauses. The choice of these clausal types was inspired by a classification of clauses which Kameyama (1998) adopts in an extended version of *Centering*.

The semantic types we consider in the following are: *property*, *eventuality*, *fact-like object*, *proposition-like object*. The latter three types are taken from the middle layer of the hierarchy of saturated abstract objects proposed by Asher (1993). The type *eventuality* comprises the types *state* and *event* which includes *activity*, *process*, *accomplishment* and *achievement*. *Fact-like object* includes *possibility*, *situation*, *fact* and *state of affairs*, while *proposition-like object* comprises *pure proposition*, *question*, *command* and *desire*. For simplicity in this paper we have included referents coded as *speech act* in the *proposition-like object* type.

Referent type	**Det**	**Dette**	**Total**
CL			
eventuality	15	2	17
fact-like	13	17	30
proposition-like	5	1	6
total	33	20	53
CCL			
fact-like	4	7	11
proposition-like	1	1	2
total	5	8	13
SCL			
eventuality	16	11	27
fact-like	14	17	31
proposition-like	6	4	10
total	36	32	68
MCL			
eventuality	2	0	2
fact-like	1	0	1
total	3	0	3
DS			
eventuality	2	1	3
fact-like	3	5	8
proposition-like	1	1	2
total	6	7	13
VP			
eventuality	17	3	20
CPR			
property	13	1	14

Table 1: Pronouns, antecedent and referent types in texts

4 Investigating the data

We have extracted from the annotated corpora the pronominal types, their antecedent and referent types and the anaphoric distance. Differing from Navarretta (2010) we do not consider here abstract anaphora chains, that is abstract anaphors having abstract pronominal anaphors as antecedents.

4.1 Syntactic antecedent and semantic referent types

Table 1 contains the occurrences of personal and demonstrative pronouns, the syntactic type of their antecedents and the semantic type of their referents in the Danish texts.

These data indicate that the majority of the abstract anaphors in the texts had a subordinate clause or a simple main clause antecedent. The ambiguous pronoun *det* is the most common abstract anaphor with all antecedents, but especially with copula predicate or verbal phrase antecedents. Both personal and demonstrative pronouns occur with all types of referent, but the demonstrative pronoun *dette* only seldom refers to a property. Demonstrative pronouns refer more frequently to *fact-like objects* than to other types of objects.

Table 2 shows the abstract anaphors, the antecedent and referent types which occurred in the monologues.

Referent type	**Unstressed pronoun**	**Stressed pronoun**	**Total**
CL			
eventuality	1	1	2
fact-like	0	1	1
proposition-like	3	9	12
total	4	11	15
CCL			
fact-like	2	0	2
proposition-like	8	1	9
total	10	1	11
SCL			
proposition-like	1	0	1
DS			
proposition-like	1	0	1
VP			
eventuality	1	2	3
CPR			
property	3	3	6

Table 2: Pronoun, antecedent and referent types in monologues

The data in the table indicate that the most common antecedent types of abstract anaphors in the monologues are *simple main clause* and *complex clause* and that the most common referent type of the unstressed pronoun *det* is *proposition-like object*. The occurrences of pronominal, antecedent and referent types in the dialogues are in Table 3.

Referent type	Unstressed pronoun	Stressed pronoun	Total
CL			
eventuality	26	26	52
fact-like	51	44	95
proposition-like	35	7	42
total	112	77	189
CCL			
eventuality	6	6	12
fact-like	11	14	25
proposition-like	12	5	17
total	29	25	54
SCL			
eventuality	8	4	12
fact-like	7	10	17
proposition-like	6	7	13
total	21	21	42
MCL			
fact-like	1	0	1
proposition-like	0	1	1
total	1	1	2
DS			
eventuality	4	2	6
fact-like	9	4	13
proposition-like	4	4	8
total	17	10	27
VP			
eventuality	52	35	87
CPR			
property	21	11	33

Table 3: Pronouns, antecedent and referent types in dialogues

Also in the dialogues the most frequently occurring antecedent syntactic type of the abstract anaphors is *simple main clause*. *Complex clause* and *verbal phrase* are also quite frequent antecedent types. The latter type is especially frequent in the maptask dialogues. This is not surprising given that maptask dialogues are interactions between a speaker (the giver) who gives instructions to a second speaker (the follower) on how to reach a place on a map. The two speakers cannot see each other and they have two slightly different maps.

As in the other types of the DAD data also in the monologues the personal pronoun, the unstressed *det*, is the most frequently occurring abstract pronoun. The demonstrative pronouns refer more frequently to *fact-like* objects than to other types of abstract objects, while proposition-like objects are more often referred to by the personal pronouns than by the demonstrative ones, as also described by Navarretta (2010).

Concluding most abstract anaphors in our texts have a single subordinate clause or simple main clause antecedent, while in dialogues they have a simple main clause

or a complex clause antecedent. Clauses are the antecedents of both demonstrative pronouns and personal pronouns as also reported in Navarretta (2010). The analysis of the referents' semantic types also confirms the results in Navarretta (2010) and shows that proposition-like objects are more often referred to by personal pronouns than by demonstrative pronouns and that the preferred referent type of demonstrative pronouns is *fact-like object*. Matrix clauses are only seldom the antecedents of abstract anaphora.

4.2 Anaphoric distance

In Table 4 we show the distance between abstract anaphors and their antecedents in terms of clauses.

Distance	Texts	Monologues	Dialogues	Total
zero	163	34	266	463
one	13	2	119	134
two	3	1	28	32
three	2	0	13	15
four	1	0	3	4
five	1	0	0	1
six	1	0	0	1
nine	0	0	1	1
ten	0	0	1	1
eleven	0	0	2	2

Table 4: Anaphoric distance in clauses

Table 5 contains the most frequent combinations (more than 8 occurrences) of referent type, antecedent syntactic type, pronominal type and a certain anaphoric distance in the texts. Table 6 shows the same data for the dialogues and the monologues.

Total	Distance	Referent type	Antecedent type	Pronominal type
16	ZERO	FACT-LIKE	CL	dette
14	ZERO	EVENTUALITY	SCL	det
13	ZERO	EVENTUALITY	CL	det
13	ZERO	FACT-LIKE	SCL	dette
12	ZERO	EVENTUALITY	VP	det
12	ZERO	FACT-LIKE	CL	det
11	ZERO	FACT-LIKE	SCL	det
10	ZERO	EVENTUALITY	SCL	dette

Table 5: Anaphoric distance, referent, antecedent and pronominal types in texts

The data in these tables indicate that the majority of the antecedents of abstract anaphors (70.8% of the cases) in our corpora occur in the clause or are the clause which immediately precedes the clause where the anaphor occurs. In 20.5% of the cases the anaphoric distance is one, in 4.9% of the cases there are two clauses in between the anaphor and the antecedents, in 2.3% of the cases the anaphoric distance is of three clauses and in 0.6% of the cases there are four clauses in-between the anaphor and its antecedent. Larger anaphoric distance (up to 11 clauses/utterances) occur very

Total	Distance	Referent type	Antecedent type	Pronominal type
Dialogues				
30	ZERO	FACT-LIKE	CL	unstressed
28	ZERO	EVENTUALITY	VP	unstressed
27	ZERO	EVENTUALITY	VP	stressed
27	ZERO	FACT-LIKE	CL	stressed
21	ZERO	PROP-L	CL	unstressed
17	ONE	EVENTUALITY	VP	unstressed
16	ZERO	EVENTUALITY	CL	stressed
16	ONE	FACT-LIKE	CL	unstressed
13	ZERO	EVENTUALITY	CL	unstressed
10	ZERO	PROP	CPR	unstressed
10	ONE	EVENTUALITY	CL	unstressed
9	ZERO	FACT-LIKE	CCL	stressed
9	ZERO	PROP-L	CCL	unstressed
Monologues				
9	ZERO	EVENTUALITY	CL	unstressed

Table 6: Anaphoric distance, referent, antecedent and pronominal types in dialogues and monologues

seldom and only in the multiparty dialogues. It must be noticed that in dialogues short utterances like *yes* and *okay* are counted as clauses and there might occur more of them uttered by different speakers. This partly explains the occurrences of long distance abstract anaphors in our data. The syntactic types which occurred more frequently when the anaphoric distance is more than one clause are subordinate clause and discourse segments in the texts, simple main clauses and discourse segments in the dialogues. Only referents of the types *eventuality* and *fact-like objects* occur when the anaphoric distance is of more than one clause and in nearly all cases the anaphors with distant antecedents are occurrences of the unstressed pronoun in dialogues and the ambiguous pronoun *det* in texts.

5 Conclusion

In this paper we have presented an investigation of the relation between syntactic antecedent types, semantic referent types, pronominal types and anaphoric distance in Danish written and spoken data. The results of our study show that the most frequently occurring antecedent types of both abstract personal and demonstrative pronouns in texts are *subordinate clause* and *simple main clause*, while in spoken data they are *simple main clause* and *complex clause*. In the maptask dialogues *verbal phrase* is also a common antecedent type and this is not surprising given the interaction type. Matrix clauses occur only seldom as the antecedents of abstract anaphors.

The investigation of the semantic types of the referents of abstract anaphors in the data confirms the studies reported in Navarretta (2010) showing that proposition-like

objects are more often referred to by personal pronouns than by demonstrative pronouns and that the preferred referents of demonstrative pronouns are fact-like objects.

The study of the anaphoric distance indicates that 70.8% of the abstract anaphors occur in the clause that immediately follows the clause in which the antecedent is (or the clause that is the antecedent) . In 20.5% of the cases there is a clause in between the anaphor and the antecedent and in 4.9% of the cases there are two clauses in between the anaphor and its antecedent. Larger anaphoric distance occur seldom and mostly in the dialogues. The syntactic types which occurred most frequently at long distance are subordinate clause and discourse segments in the texts, simple main clauses and discourse segments in the dialogues. Only eventuality and fact-like referents are referred to when the anaphoric distance is of more than one clause and in nearly all cases the anaphor with distant antecedents is the unstressed pronoun in dialogues and the ambiguous pronoun *det* in texts.

Currently we are testing the relation between the various annotation types using machine learning algorithms.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366, 1991.

Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87, 1988.

Mira Ariel. Interpreting anaphoric expressions: a cognitive versus a pragmatic approach. *Journal of Linguistics*, 30(1):3–40, 1994.

Nicholas Asher. *Reference to Abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

Kaja Borthen, Thorstein Fretheim, and Jeanette K. Gundel. What brings a higher-order entity into focus of attention? Sentential pronouns in English and Norwegian. In Ruslan Mitkov and Branimir Boguraev, editors, *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 88–93, 1997.

Donna K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 80–87, 2002.

Donna K. Byron and James Allen. Resolving demonstrative pronouns in the TRAINS93 corpus. In *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC 1998)*, pages 68–81, 1998.

Miriam Eckert and Michael Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89, 2001.

Kari Fraurud. *Processing Noun Phrases in Natural Discourse*. Department of Linguistics - Stockholm University, 1992.

Frans Gregersen. The LANCHART Corpus of Spoken Danish, Report from a corpus in progress. In *Current Trends in Research on Spoken Language in the Nordic Countries*, pages 130–143. Oulu University Press, 2007.

Nina Grønnum. DanPASS - A Danish Phonetically Annotated Spontaneous Speech Corpus. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genova, Italy, May 2006.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12:281–299, 2003.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Demonstrative pronouns in natural discourse. In A. Branco, T. McEnery, and R. Mitkov, editors, *Proceedings of DAARC-2004 - The 5th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 81–86, Furnal, S.Miguel, Portugal, 2004. Ediçoes Colibri.

Nancy Hedberg, Jeanette K. Gundel, and Ron Zacharski. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In António Branco, T. McEnery, Ruslan Mitkov, and F. Silva, editors, *Proceedings of DAARC-2007 - the 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 31–36, Lagos, Portugal, March 2007.

Michael Hegarty. Semantic types of abstract entities. *Lingua*, 113:891–927, 2003.

Elsi Kaiser. Pronouns and demonstratives in Finnish: Indicators of Referent Salience. In P. Baker, A. Hardie, T. McEnery, and A. Siewierska, editors, *Proceedings of the Discourse Anaphora and Anaphor Resolution Conference*, volume 12 of *University Center for Computer Corpus Research on Language - Technical Series*, pages 20–27, Lancaster, UK, 2000.

Megumi Kameyama. Intrasentential centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Oxford University Press, Oxford, U.K., 1998.

Britt Keson and Ole Norling-Christensen. PAROLE-DK. Technical report, Det Danske Sprog- og Litteraturselskab, http://korpus.dsl.dk/e-resurser/parole-korpus.php, 1998.

Stephen C. Levinson. Pragmatics and the grammar of anaphora: A partial pragmatic reduction of Binding and Control Phenomena. *Journal of Linguistics*, 23(2):379–434, 1987.

John Lyons. *Semantics*, volume I-II. Cambridge University Press, 1977.

Christoph Müller. Resolving It, This and That in unrestricted multi-party dialog. In *Proceedings of ACL-2007*, pages 816–823, Prague, 2007.

Costanza Navarretta. Centering-based anaphora resolution in Danish dialogues. In P. Sojka, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue - Proceedings of the 3rd International Workshop (TSD 2000)*, pages 345–350, Brno, Czech Republic, 2000.

Costanza Navarretta. *The Use and Resolution of Intersentential Pronominal Anaphora in Danish Discourse*. PhD thesis, Centre of Language Technology and Department of General and Applied Linguistics Copenhagen University, 2002.

Costanza Navarretta. The main reference mechanisms of Danish demonstrative pronominal anaphors. In A. Branco, T. McEnery, and R. Mitkov, editors, *Proceedings of DAARC-2004 - the 5th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 115–120, Furnal, S.Miguel, Portugal, 2004. Ediçoes Colibri.

Costanza Navarretta. Resolving individual and abstract anaphora in texts and dialogues. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 233–239, Geneva, Switzerland, 2004a.

Costanza Navarretta. A contrastive analysis of the use of abstract anaphora. In A. Branco, T. McEnery, R. Mitkov, and F. Silva, editors, *In Proceedings of DAARC-2007 - 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 103–109, Lagos, Portugal, March 2007. Centro de Linguistica da Universidade do Porto.

Costanza Navarretta. Co-referential chains and discourse topic shifts in parallel and comparable corpora. *Revista de Procesamiento de Lenguaje Natural - La Sociedad Espaõla para el Procesamiento del Lenguaje Natural (SEPLN)*, 42:105–102, 2009.

Costanza Navarretta. The DAD parallel corpora and their uses. In *Proceedings of LREC 2010*, pages 705–712, Malta, May 2010. ELRA.

Costanza Navarretta and Sussi Olsen. Annotating abstract pronominal anaphora in the DAD project.

In *Proceedings of LREC-2008*, Marrakesh, Morocco, May 2008. ELRA.

Costanza Navarretta and Sussi Olsen. The annotation of pronominal abstract anaphora in Danish texts and dialogues. DAD report 1, Centre for Language Technology, University of Copenhagen, January 2009.

Luigi Pirandello. *Novelle per un anno*. Giunti, 1922-1937.

Michael Strube and Christoph Müller. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the ACL'03*, pages 168–175, 2003.

Bonnie L. Webber. Discourse deixis and discourse processing. Technical report, University of Pennsylvania, 1988.

Bonnie L. Webber. Structure and ostension in the interpretation of discourse deixis. *Natural Language and Cognitive Processes*, 6(2):107–135, January 1991.

Information Structure Annotation and Secondary Accents

Arndt Riester[1] *& Stefan Baumann*[2]
[1]IMS, University of Stuttgart
[2]IfL Phonetics, University of Cologne

Abstract
We present a proposal for an annotation system for information structure that combines contemporary corpus-oriented accounts of information status with insights from the recent theoretical debate (e.g. Selkirk, 2007; Beaver & Velleman, subm.) on the basic pragmatic sources which lead to primary and secondary accentuation; in particular, the combination of the given-new distinction with the classical triggers for F-marking by Rooth (1992). We comment on the yet undecided question whether one or several kinds of focus should be considered in the annotation task. A key property of our scheme is its distinction between a lexical and a referential level. This allows us to describe fine-grained properties of texts, e.g. the information structurally and prosodically relevant observation that a given discourse referent may be taken up by means of lexically new material. The annotation system is demonstrated for examples from transcripts of spoken corpora as well as sentences taken from the theoretically oriented literature. We report on the inter-annotator agreement reached, and show how the system can be used in the investigation of subtle prosodic phenomena like secondary accents, which have been claimed to mark second occurrence focus.

1 Contrastive focus vs. information focus

A longstanding issue in information structure theory is the differentiation between so-called *contrastive focus* and *information focus* (focus related to the novelty of a constituent). Both types of focus are commonly marked by primary pitch accents, i.e. by strong prosodic prominence. While the distinction is usually demonstrated on the basis of intuitive minimal pairs like (1), from Selkirk (2007), its fundamental semantic distinction has remained controversial.

(1) a. I gave one to SArah$_{CF}$, not to CAITlin$_{CF}$.
b. I gave one to SArah$_{IF}$.

What examples like (1) seem to suggest is that *contrastive focus* requires the overt availability of a pair of alternatives. One problem of contemporary focus literature is that, usually, cases like (1a) are grouped together with examples involving focus-sensitive particles like (2a) or question-answer sequences like (2b), following the paradigm of Rooth (1992).

(2) a. Semanticists only talk about ONly$_F$.
b. What did the semanticist talk about this time? She talked about ANSwers$_F$.

No overt alternatives are involved in these examples, which has led researchers to quite different reactions. An obvious move is to abandon the notion of *contrastive focus* altogether, and to try and find a uniform explanation for focusing in general. This is the path that is pursued by the "lumpers" camp[1], e.g. Büring (to appear); Rooth

[1]Quote: Beaver and Velleman (subm., Sect. 1)

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 111-127, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

(2010). By contrast, "splitters" like Selkirk (2007) and Beaver and Velleman (subm.) explicitly claim that it is necessary to distinguish between two *sources* that determine the assignment of *prominence* – while avoiding the question whether there are one or several *types* of focus. The two sources are, on the one hand, *novelty* (marked in logical form by means of an N feature or, indirectly, by lack of a *G(iven)* feature) and, on the other hand, a collection of factors which are varyingly pooled under the notions of either "contrastive focus" (Götze et al., 2007), "focus" (Rooth, 1992; Selkirk, 2007) or "importance" (Beaver and Velleman, subm.). They are usually assumed to carry an F feature in logical form and to share the common property of evoking a set of (implicit or explicit) *alternatives*.

Since the terminological situation is obviously complicated we try to be careful in using notions like "focus". We acknowledge the need for a two-factor account of identifying the information structural setup of sentences and discourses. However, we think that we would go too far if we reserved the *focus* notion for expressions whose prominence is due to an alternative-related property such as explicit contrast, the presence of exhaustive particles or a wh-question. If consequently applied, this would lead to the conclusion that some standard examples of focus such as (3a,b) no longer ARE focus examples since the only obvious reason for the accents at hand is the novelty of their host phrases.

(3) a. Mary went into a store. She [bought a book about BATS].
b. Let me tell you a secret about Sally and John. Sally is [in LOVE] with John.

This, however, is the situation that we permanently encounter in transcripts from monologues and many other kinds of texts. A differentiation between answers to overt questions and ordinary, unsolicited, information-conveying sentences is artificial also for the reason that there are theories explaining the structuring of discourse by use of (often implicit) *questions under discussion*, e.g. Roberts (1996).

The position we are mildly favouring is, therefore, to not exclude the use of the term *focus* in cases like (3a,b), which only involve given and new information, and to use the term *elicited alternatives* for direct sources of focusing (F-marking). The more fine-grained a classification system is, the less does the question matter whether the classes are reducible to one or two types of focus. Instead, we would like to find out whether these fine-grained differences that we can detect at the pragmatic level are related to subtle differences at the prosodic level. We are not only interested in the location and realisation of the nuclear accent but also in pre- and postnuclear secondary accents and other prosodic phenomena.

2 Annotation of information structure

In this paper, we make a proposal for the annotation of *information structure*. We use this rather underspecified term to avoid the intricacies, discussed in the previous

section, surrounding the notion of *focus*. Moreover, we take notice of the persisting terminological confusion in the field, an unfortunate matter which we do not expect to be overcome soon. However, we feel the need to clarify which aspects of information structure we are interested in. Our main interest lies in the focus-background distinction, which we are going to analyze in more detail than usually seen in contemporary accounts. Our basic coordinates are the formal accounts of focus as provided, on the one hand, by Rooth (1992), and on the other hand, by Schwarzschild (1999), which we take to describe complementary, yet compatible, aspects of the focus notion.[2]

According to other accounts, *topic* is taken to be the complement of focus (Hajičová et al., 1998). However, we follow e.g. Krifka (2007) in distinguishing the topic-comment dimension from the focus-background dimension. We will not be discussing the former. Neither shall we consider a theme-rheme distinction.

2.1 Previous approaches to annotating information structure

Annotations of focus and related information structural features are often said to be "difficult" as compared to, for instance, morphosyntactic annotations. This is likely due to the fact that informal definitions of focus are often remarkably vague whereas insights from the formal-semantic literature are not easily transfered to corpus data.

For instance, in her study of focus and topic in a corpus of spoken Danish, Paggio (2006) defines focus, quoting Lambrecht (1994), as "non-presupposed information". This, in combination with a number of heuristics and general principles (such as "all sentences have a focus") is used as a guideline for the annotators. Paggio reports a kappa score between 0.7 and 0.8 on controlled monologue and dialogue data such as descriptions and map tasks. In her setting, however, annotators made use of prosodic information, which makes the annotation task simpler but also semantically intransparent.

In the *LISA* (Linguistic Information Structure Annotation) guidelines (Götze et al., 2007), information structure is annotated on three layers: information status (*given / accessible / new*, restricted to referring expressions), topic and focus. Focus is defined as "[t]hat part of an expression which provides the most relevant information in a particular context" (p.170). "New-information focus" is distinguished from "contrastive focus". *New information* may come as *solicited* (in response to a question) or *unsolicited*.[3] The guidelines additionally contain a useful list of triggering constellations for contrastive focus (see Sect. 3.3 below). As for the focus layer, Ritz et al. (2008) report an inter-annotator agreement of 0.41 to 0.62 for different types of texts based on predefined markables (but no prosodic information). Telling from these rather low scores, annotating focus has not yet reached a satisfactory level.

[2] It has been noted in Beaver and Clark (2008, Sect. 2.4), though, that the usage of F-features is not the same on the two accounts.

[3] Note that this stands in contrast to e.g. Rooth (1992); Selkirk (2007), who would count solicited, but not unsolicited, information as a trigger for F-marking akin to contrastive focus.

We specifically want to point out what we see as an unfortunate decision in the LISA guidelines: the choice to separate the annotations of, on the one hand, information status, and, on the other hand, new information focus. Both describe the given-new distinction, thus, the same kind of information. They differ in that information status allows for a more differentiated classification but is limited to referential expressions. It is our explicit goal to overcome this separation and, thereby, generalise the notion of information status to all expression types.

2.2 A new labeling system for information status: the RefLex scheme

In the following, we will briefly introduce a labeling system for information status, which distinguishes between a referential level and a lexical level (and which we therefore call the *RefLex* scheme). We will clarify why it is desirable to use such a fine-grained system rather than just distinguishing between "given" and "new" constituents. Note that we are not claiming that the annotation labels presented below represent syntactic features of some kind, in the way as, for instance, Selkirk (2007) treats her F and G markings. We will make no predictions as regards the precise functioning of the syntax-phonology interface. Nevertheless, a crucial point of the whole procedure is the assumption that the information status labels have an impact on prosody.

The category descriptions below are kept very short, since we have introduced them in great detail elsewhere (Baumann and Riester, subm.). By use of the following choice of R-categories it is possible to classify *referential* determiner phrases and prepositional phrases occurring in natural discourse; by use of the L-categories we can classify the information status of content words and non-referential phrases.

2.2.1 R-GIVEN and L-GIVEN

Givenness, loosely following Schwarzschild (1999, 151), can be interpreted as either synonymy / hyponymy of lexemes or as identity between referring expressions. Likewise, Halliday and colleagues[4] distinguish between *lexical cohesion* and various referential relations. We call the two notions *L-givenness* and *R-givenness*, respectively. Interesting constellations can be observed if the two notions are simultaneously applied, as shown below.

R-labels apply at the DP or PP level. For instance, in examples (4), (5) and (7) we find various kinds of coreferential expressions. Lexical givenness, on the other hand, applies in (5) and (7) on the repeated words, and in (6) on the hypernym "guy".

(4)

A colleague came in.	The	idiot	dropped a vase.
	R-GIVEN		

(5)

A student came in.	Another	student	greeted	him.
		L-GIVEN		
				R-GIVEN

[4] e.g. Halliday and Hasan (1976, 288); Halliday and Matthiessen (2004)

(6)

A policeman came in.	Another	guy	left.
		L-GIVEN	

(7)

A man came in.	The	man	coughed.
		L-GIVEN	
	R-GIVEN		

The most important take-home message is that neither is referential givenness a prerequisite for lexical givenness, as shown in (4), nor the other way round, see (5) and (6), although the two sometimes combine, as in (7).

2.2.2 R-NEW, L-NEW, R-UNUSED

Novelty is, on most treatments of information structure and discussions of the *given/new* distinction, understood as "novelty in the discourse". Remarkably, however, Prince (1992) additionally distinguishes between *discourse novelty* and *hearer novelty*, the latter representing a stronger notion since unmentioned (i.e. discourse-new) entities may nevertheless be familiar to the addressee (i.e. hearer-old). In her earlier paper, Prince (1981) uses the labels *unused* (discourse-new, hearer-old) and *brand-new* (discourse-new, hearer-new) for the same opposition. The labels R-NEW and R-UNUSED that are employed on our account are defined in a slightly different way: both describe discourse-new referential expressions but, while R-NEW is reserved for indefinites, R-UNUSED stands for uniquely identifiable, definite, but not necessarily *known*, entities used on the first occasion in a text. This decision, on the one hand, does justice to the long-standing semantic tradition to keep indefinites and definites (for instance, proper names) apart, and, on the other hand, accounts for the difficulty to decide with certainty whether, for instance, a *named entity* is hearer-known or not[5].

Independently of what has just been said, it is furthermore possible to separately describe the discourse novelty of *lexemes* (L-NEW) and of the *discourse referents* (R-NEW, R-UNUSED) which they introduce. Examples of the three categories in combination are given in (8) to (10).

(8)

A	man	came in.	Another	man	left.
	L-NEW			L-GIVEN	
R-NEW			R-NEW		

(9)

George	came in.	Mary	likes	George.
L-NEW		L-NEW		L-GIVEN
R-UNUSED		R-UNUSED		R-GIVEN

(10)

The	man	who stole	my	wallet	is very tall.
	L-NEW			L-NEW	
			R-UNUSED		
R-UNUSED					

[5] Although the respective subclassifications can be made with a reasonable degree of agreement, cf. Riester et al. (2010).

2.2.3 R-BRIDGING, L-ACCESSIBLE

Prince (1981) and also Chafe (1994) have pointed out that it is desirable to not only distinguish between *given* and *new* information but to take into account at least a third, intermediate, class: expressions which have not been mentioned explicitly but are *inferrable* from material in the discourse. Chafe (1994) uses the term *accessible* for such information but he does not distinguish between different levels, as we would like to do. As far as referents are concerned, a closely related phenomenon has been discussed under the notion of *bridging* (Clark, 1977; Asher and Lascarides, 1998), shown in example (11).

(11)

Bill	discovered	a romantic	house.	The	door	was open.
L-NEW			L-NEW		L-ACCESSIBLE	
R-UNUSED		R-NEW		R-BRIDGING		

The label L-ACCESSIBLE is defined for words which are hyponyms or meronyms (part expressions) of other words in the recent discourse context (i.e. not further away than 5 clauses). The label R-BRIDGING, on the other hand, is defined quite differently as a definite expression whose licensing depends on a previously introduced scenario or frame. So, while in (11), "door" and "house" stand in a part-whole relation ("door" is lexically accessible), no such relation exists between "murdered" and "harpoon" in (12). Since the harpoon is an unusual murder instrument, it is labeled L-NEW. Nevertheless, we would still like to say that this is a case of bridging, since the second sentence could not be uttered felicitously at the beginning of a discourse.

(12)

John	was murdered yesterday.	The	harpoon	was lying nearby.
L-NEW			L-NEW	
R-UNUSED		R-BRIDGING		

Other than R-UNUSED expressions, the interpretation of items labeled R-BRIDGING is context-dependent. In contrast to the label R-GIVEN, R-BRIDGING implies non-coreference. Indefinites never receive the label R-BRIDGING in the present system. In (13), lexical accessibility combines with referential novelty.

(13)

John	lives	in	Italy	and is married	to a	Neapolitan.
L-NEW			L-NEW			L-ACCESSIBLE
R-UNUSED		R-UNUSED			R-NEW	

Note that identifying R-BRIDGING as a separate referential class of information status in between given and new expressions does not only derive from purely theoretical considerations but can be shown to have a significant influence on the realisation of nuclear accents; an issue which is highly relevant for information structure theory. Röhr and Baumann (2010) demonstrate in experiments that *inferrable* information is significantly more often produced with low or falling accents as compared to new information, which is predominantly realised with (perceptually more prominent) high accents. An

example is shown in (14), whose prosodic realisation (with an early peak accent that is low on the accented syllable) can be seen in Figure 1.[6]

(14) Thomas darf heute im Zoo seinen Lieblingsaffen füttern. [...] Er steckt sich [R-BRIDGING die [L-NEW Banane]] ein.
'Today, Thomas has got the permission to feed his favourite monkey at the zoo. He pockets the banana.'

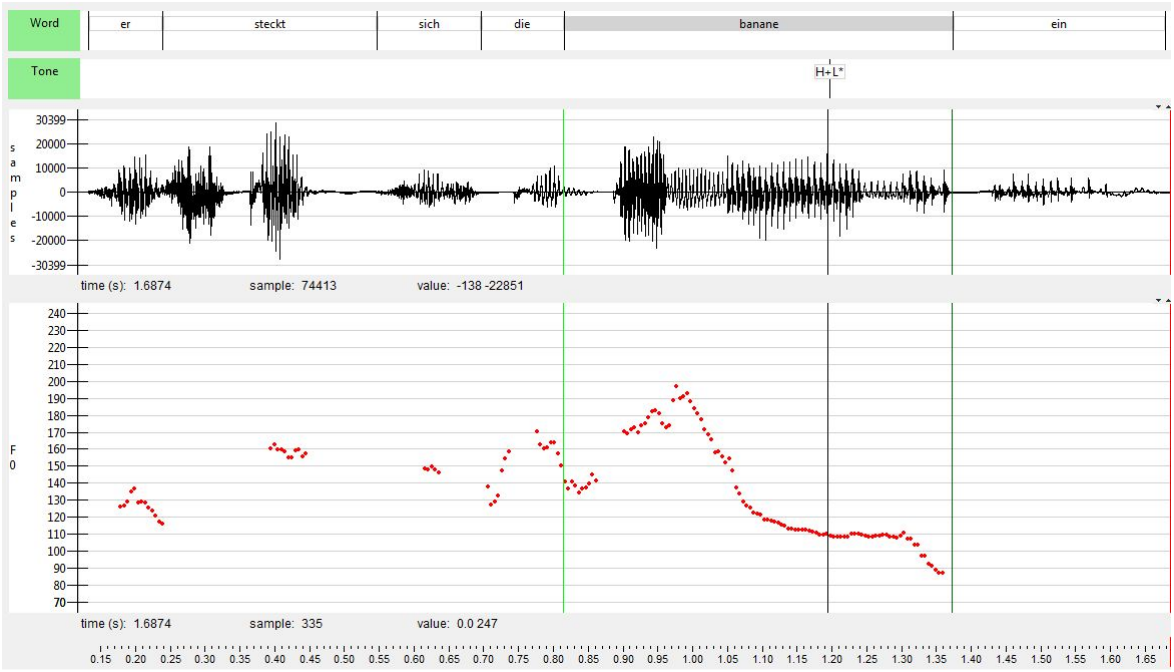


Figure 1: Possible realisation of an L-NEW, R-BRIDGING expression ("die Banane")

2.2.4 R-GENERIC

Definite or indefinite expressions which refer to a kind, see (15) and (16), receive the label R-GENERIC.

(15)

The	fox	is	a	predator.
	L-NEW			L-NEW
R-GENERIC			R-GENERIC	

(16)

Mary	only likes	vegetables.
L-NEW		L-NEW
R-UNUSED		R-GENERIC

The examples of R and L labels presented in the sections above only show a small number of combinations that are possible in the annotation system, which allows for very detailed information structural investigations of discourses. For a comprehensive list of possible combinations consult Baumann and Riester (subm.). In the following section we will turn to a number of practical issues which arise when we apply the

[6]Screen shot of the target sentence using the speech analysis tool EMU (Cassidy and Harrington, 2001), displaying labeling tiers for words and intonation (accents annotated according to GToBI, following Grice et al. (2005), as well as the oscillogram and pitch contour)

annotation scheme to corpus data. Finally, we present some proposals for using the system in the task of identifying and describing regions of texts which are particularly interesting as far as prosody is concerned. Some of these have received wide attention in the semantic literature, such as so-called *second occurrence focus*.

3 Annotating corpus data

3.1 Annotation of syntactic phrases

In previous literature on information status (Prince, 1981, 1992; Nissim et al., 2004; Götze et al., 2007; Riester et al., 2010) usually only referential expressions (syntactically: DPs, PPs) are considered as the units for annotation. However, ever since in the development of information *structure* theory, givenness and novelty have been defined for all syntactic categories.

It is our claim that, in defining information status at the L-level, we are providing the foundations for a comprehensive information structural analysis of sentences and discourses. Of course, the question what counts as a unit for annotation is influenced by the choice of syntactic theory underlying the analysis. For the time being, we shall be classifying projections of *content words* like verbs, nouns, adjectives and adverbs, i.e. non-referential phrases of category VP, NP,[7] AP, AdvP and S. A basic overview of what counts as an R-level or an L-level unit is shown in Fig. 2.

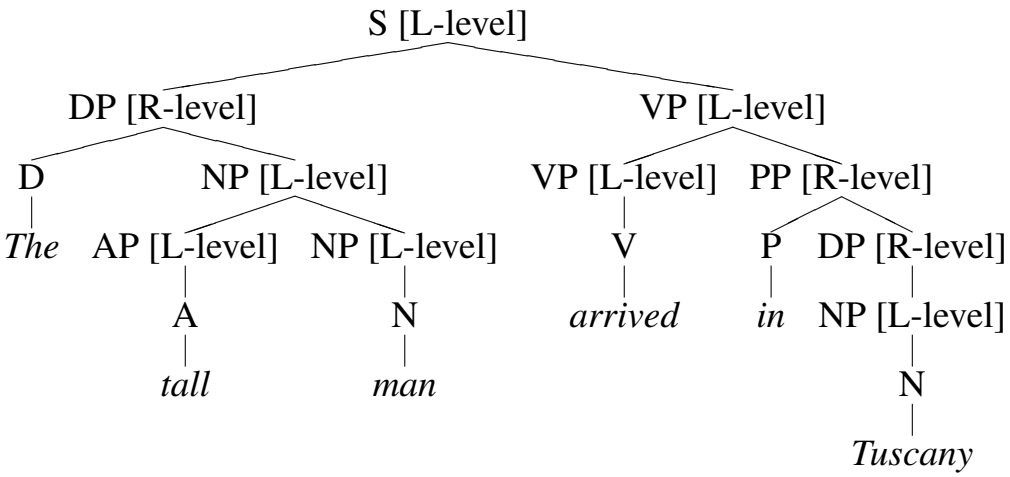


Figure 2: Basic target units for *RefLex* annotations

We would like to point out that what we are proposing amounts to a practical explication – and further development – of the approach taken by Schwarzschild (1999, 151), who distinguishes between categories of type e (R-level) and of type $\langle \alpha, \beta \rangle$ (L-level). Our definition of the L-level, however, is much simpler than Schwarzschild's since we completely relinquish his notion of *Existential F-closure*. However, we make use of

[7] We are assuming the DP hypothesis. Accordingly, we take NPs to denote properties, i.e. sets of individuals, whereas DPs denote (or refer to) a single individual or group entity.

his idea to generalise lexical relations to a notion of entailment.[8] In corpus annotation practice, the linguistic scheme shown in Fig. 2 will have to be adapted to various constraining factors, such as the properties of the chosen parser with its specific syntactic tagset, as well as features of the annotation tool. Fig. 3 shows the annotation of a German sentence, which was parsed using XLE and the German LFG grammar by Rohrer and Forst (2006), and converted to be used with the SALTO tool (Burchardt et al., 2006), which produces output in TIGER/SALSA-XML. In the rest of the paper, we shall abstract over such individual choices, since it is our goal to provide the general annotation procedure and not one that is tied to a specific annotation tool, format or syntactic theory.

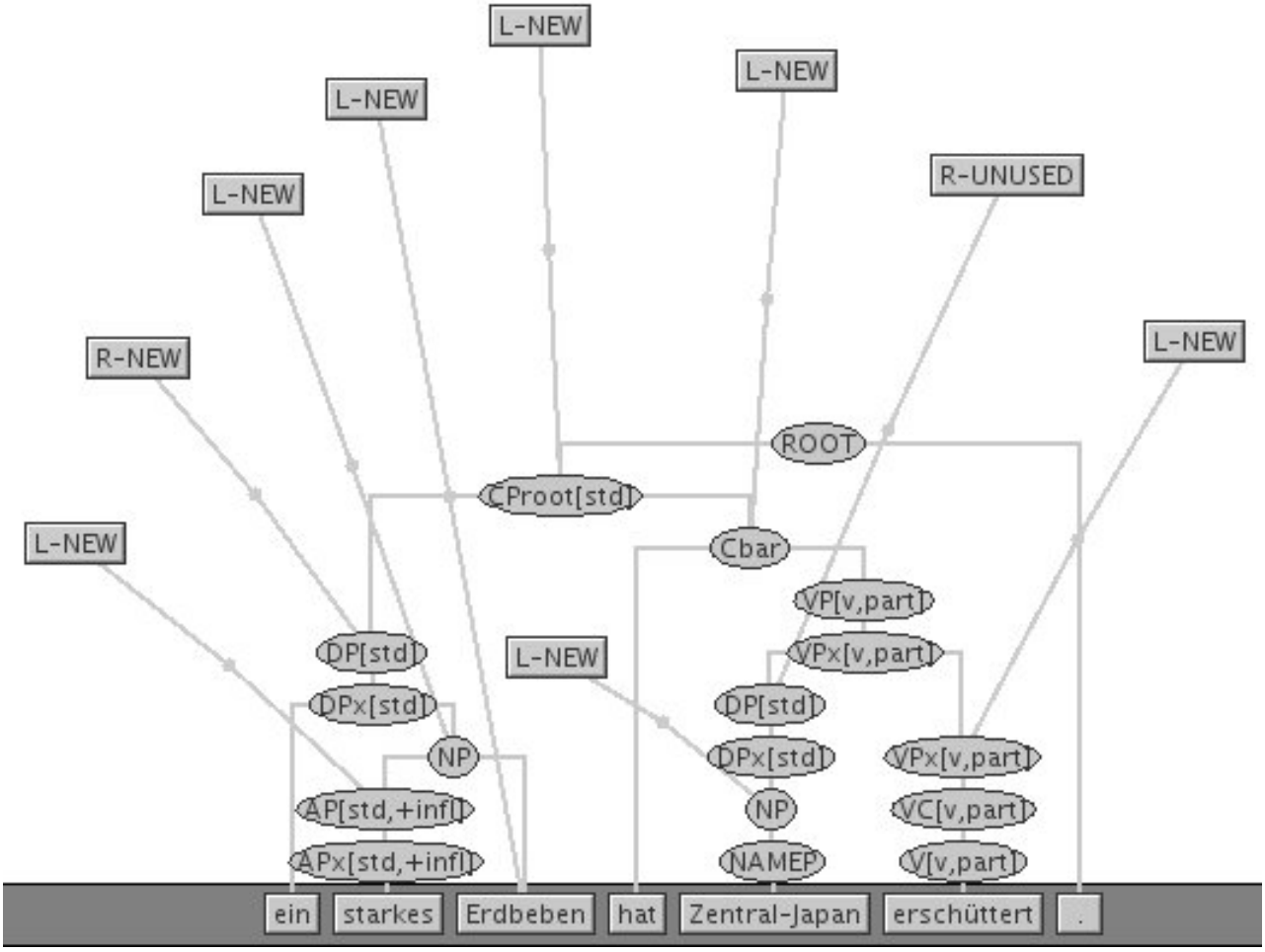


Figure 3: Sentence annotated in SALTO: *A strong earthquake has hit central Japan.*

3.2 Example annotation of a radio news feature

In the following, we will briefly show how the extended annotations can be applied to an example from a German radio news bulletin before turning to the reanalysis of some theoretically more advanced examples. The news example is (17), which will be

[8] According to this approach, the previous mention of "chihuahua" **entails** the successively mentioned hypernym "dog", as well as a successive mention of "small dog", although we wouldn't normally want to say that the latter phrase is a "hypernym", cf. Baumann and Riester (subm., Sect. 3).

annotated as in (18-20). We use a simplified table notation and additionally provide the GToBI labels (Grice et al., 2005) for the corresponding speech data.[9] Note, however, that in our envisaged annotation process, the labelers will have no access to prosodic information, since it is the correspondence between prosody and information structure which we are ultimately intending to investigate.

(17) a. Ein starkes Erdbeben hat Zentral-Japan erschüttert.
'An strong earthquake has hit central Japan.'

b. Die Behörden gaben eine Tsunami-Warnung für den Südwesten heraus.
'The authorities have issued a tsunami warning for the Southwest.'

c. Auch im Inselstaat Vanuatu im Südpazifik wurden zwei Beben registriert.
'Also in the island state of Vanuatu in the Southern Pacific two earthquakes have been registered.'

(18)

	H*	L+H*		H*	H*	H+!H* L-%
Ein	starkes	Erdbeben	hat	Zentral-	Japan	erschüttert.
	L-NEW	L-NEW		L-NEW		L-NEW
	L-NEW			R-UNUSED		
R-NEW			L-NEW			
L-NEW						

(19)[10]

	H*			L+H*		L+H*	L-%
Die	Behörden	gaben	eine	Tsunami-Warnung	für den	Südwesten	heraus.
	L-NEW			L-NEW		L-NEW	L-NEW
R-BRIDGING					R-BRIDGING		
			R-NEW				
	L-NEW						
L-NEW							

(20)

H*		H*	L+H*		L+H*		L*	L*+H	H+L* L-%
Auch	im	Inselstaat	Vanuatu	im	Südpazifik	wurden	zwei	Beben	registriert.
		L-NEW	L-NEW		L-NEW			L-GIVEN	L-NEW
				R-UNUSED			R-NEW		
	R-UNUSED								
						L-ACCESSIBLE			
L-ACCESSIBLE									

Next to the general assignment of L-labels to verbal and adjectival phrases and clauses, there are a few important observations which relate to complex phrases like [$_{\text{R-NEW}}$ eine Tsunami-Warnung [$_{\text{R-BRIDGING}}$ für den Südwesten]] in (19), or [$_{\text{R-UNUSED}}$ im Inselstaat Vanuatu [$_{\text{R-UNUSED}}$ im Südpazifik]] in (20). In each, one referential phrase has another one embedded in it. Since the two possess different referents, two R-labels are nested inside each other.

3.3 Elicited alternatives

In Sect. 1, we already discussed the need to consider two main sources that may have an influence on the prosodic realisation of a sentence: besides *information status* we

[9] One of the anonymous reviewers requested that we include the respective prosodic information. Unfortunately, at the current stage, it is impossible to provide a satisfactory discussion of the discourse-prosody interface of this example, especially since prosodic correlates of information structure usually require a broad-scale statistical analysis.

[10] The particle verb "gaben…heraus" ("issued") is annotated on "heraus".

have to identify features that are linked to Alternative Semantics. Götze et al. (2007, 178ff.) provide a number of such features under the heading of *contrastive focus*. Since we think that neither "focus" nor "contrast" are ideal labels for this class of features, for reasons discussed above, we will simply use the label ALT. A minimal list of important triggers of alternatives is shown in Table 1.

Sublabel of ALT	Description
FS	Item is associated with a focus-sensitive particle.
OC	Item is an element of a pair or list of overtly contrastive expressions (sentence-internally or across sentences); this subsumes e.g. *corrections* and *coordinated* expressions.
SE	Item selects one element from a pair or list of *previously* introduced alternatives.
VF	Verum focus

Table 1: Configurations which elicit alternatives

We think that Table 1 summarizes the relevant alternative-eliciting features. Note that, for instance, the prosodic prominence of an answer to an overt question is already adequately described by means of novelty at the R- or L-level, or the feature ALT-SE.[11]

When we apply this additional set of features to example (20), we obtain the following additional tier of *elicited alternatives* shown in (21).

(21)

<table>
<tr><td>H*</td><td></td><td>H*</td><td>L+H*</td><td></td><td>L+H*</td><td></td><td>L*</td><td>L*+H</td><td>H+L* L-%</td></tr>
<tr><td>Auch</td><td>im</td><td>Inselstaat</td><td>Vanuatu</td><td>im</td><td>Südpazifik</td><td>wurden</td><td>zwei</td><td>Beben</td><td>registriert.</td></tr>
<tr><td rowspan="4"></td><td></td><td>L-NEW</td><td>L-NEW</td><td></td><td>L-NEW</td><td>L-NEW</td><td></td><td>L-GIVEN</td><td>L-NEW</td></tr>
<tr><td colspan="2"></td><td colspan="2">R-UNUSED</td><td rowspan="2"></td><td rowspan="2"></td><td colspan="2">R-NEW</td><td rowspan="2"></td></tr>
<tr><td colspan="4">R-UNUSED</td><td colspan="2"></td></tr>
<tr><td colspan="7">L-ACCESSIBLE</td></tr>
<tr><td colspan="10">L-NEW</td></tr>
<tr><td></td><td colspan="5">ALT-FS / -OC</td><td></td><td></td><td></td><td></td></tr>
</table>

We observe that the phrase "im Inselstaat Vanuatu im Südpazifik" is associated with the additive particle "auch". It furthermore contrasts with "Zentral-Japan".

4 Inter-annotator agreement

We are now briefly going to discuss the inter-annotator agreement that we achieved for the proposed scheme, in particular for the two levels of information status. In a small experiment the two authors of this article independently annotated a text consisting of a transcript from spontaneous speech, comprising 65 sentences. Beforehand, we agreed on the set of markables to be annotated. In total, R-annotations were assigned to

[11] Likewise, we think that we do not need the feature like *implication* (Götze et al., 2007, 181), which again can be captured with our label L-NEW.

133 markables, L-annotations were assigned to 275 markables, following the schemes summarized in Table 2.[12]

R-Level		L-Level	
Units: DP, PP, *that*-CP		Units: AP, AdvP, NP, VP, S	
Label	Description	Label	Description
R-GIVEN	corefential anaphor	L-GIVEN	word identity / synonym / hypernym / holonym / superset
R-BRIDGING	non-coreferential context-dependent expression	L-ACCESSIBLE	hyponym / meronym / subset / otherwise related
R-UNUSED	definite discourse-new expression	L-NEW	unrelated expression (within last five clauses)
R-NEW	specific indefinite		
R-GENERIC	generic definite or indefinite		
OTHER	e.g. cataphors		

Table 2: Overview basic *RefLex* scheme

We achieve a κ score (Cohen, 1960) of 0.70 for the R-level and 0.78 for the L-level. We were not able to provide results for the annotation of elicited alternatives since the text chosen contained only 9 markables for ALT-labels.

5 Second occurrence focus and secondary accents

In the remaining part of this article we turn to an issue which has received much attention in both the theoretical and experimental literature: second occurrence focus, see example (22) from Partee (1999). We discuss this phenomenon in order to show how our annotation scheme for information structure ultimately might pave the way to a corpus analysis of second occurrence focus and other phenomena involving secondary (i.e. weaker) accents.

(22) A: Everyone already knew that Mary only eats VEgetables$_F$.
B: If even PAUL knew that Mary only eats VEgetables$_{SOF}$ then he should have suggested a different REStaurant.

Describing the precise conditions which license second occurrence focus (SOF) is not straightforward. Selkirk (2007) characterizes a *SOF* as a given constituent (since it has been mentioned before) which is at the same time focused (in (22) due to association

[12] See Baumann and Riester (subm., Sect. 4) for an extended scheme.

with "only") and whose antecedent is also focused. Beaver and Velleman (subm.) avoid reference to focusing by saying that a *SOF* must be "important" (F-marked, see above) as well as "predictable" (roughly: part of a larger constituent which is also given). Following our proposed annotation scheme, example (22) will receive the analysis given in (23-24).

(23)

<table>
<tr><td>Everyone</td><td>already</td><td>knew</td><td>that</td><td>Mary</td><td>only</td><td>eats</td><td>vegetables.</td></tr>
<tr><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4">L-NEW</td><td rowspan="4"></td><td>L-NEW</td><td rowspan="3"></td><td rowspan="2">L-NEW</td><td>L-NEW</td></tr>
<tr><td>R-UNUSED</td><td>R-GENERIC</td></tr>
<tr><td colspan="2">L-NEW</td></tr>
<tr><td colspan="5">L-NEW</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>ALT-FS</td></tr>
</table>

(24)

<table>
<tr><td>If</td><td>even</td><td>Paul</td><td>knew</td><td>that</td><td>Mary</td><td>only</td><td>eats</td><td>vegetables</td><td>then ...</td></tr>
<tr><td rowspan="5"></td><td rowspan="5"></td><td>L-NEW</td><td rowspan="5">L-GIVEN</td><td rowspan="5"></td><td>L-GIVEN</td><td rowspan="4"></td><td rowspan="2">L-GIVEN</td><td>L-GIVEN</td><td rowspan="5"></td></tr>
<tr><td rowspan="4">R-UNUSED</td><td>R-GIVEN</td><td>R-GENERIC</td></tr>
<tr><td colspan="2">L-GIVEN</td></tr>
<tr><td colspan="3">L-GIVEN</td></tr>
<tr><td colspan="4">R-GIVEN</td></tr>
<tr><td></td><td></td><td>ALT-FS</td><td></td><td></td><td></td><td></td><td></td><td>ALT-FS</td><td></td></tr>
</table>

Beaver et al. (2007) showed that the word "vegetables" in (24) is realised with a secondary accent which is not marked by pitch movement but rather by means of increased duration of the focused word in comparison with a deaccented version.

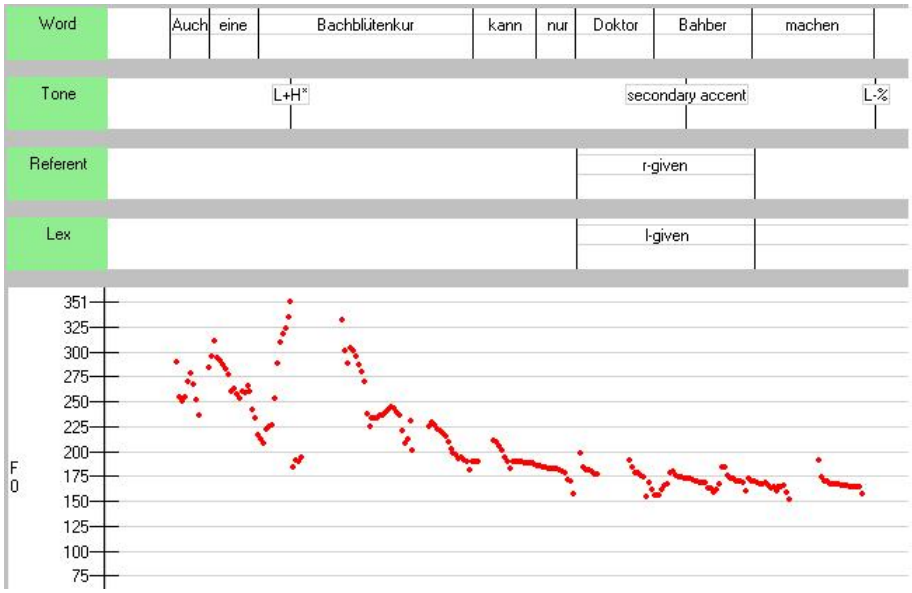


Figure 4: Realisation of second occurrence focus (R-GIVEN, ALT-FS) in German ("Dr. Bahber")

Similar results were found for German by Ishihara and Féry (2006) as well as Baumann et al. (2010). Figure 4 shows an example of second occurrence focus in German taken from the discourse in (25). The nuclear accent in (25b) is clearly placed on "Bachblütenkur", whereas "Bahber" only receives a secondary prominence.

(25) a. Eine Akupunktur kann nur Dr. Bahber machen.
'An acupuncture can only be done by Dr Bahber.'
b. Auch eine Bachblütenkur kann nur Dr. Bahber machen.'
'Also a cure with Bach flowers can only be done by Dr Bahber.' (Baumann et al., 2010, 63)

While second occurrence focus has received much attention in the literature on information structure, it is not easy to find good corresponding examples in corpus data. Nevertheless, secondary accents occur quite frequently, and it is instructive to investigate what other instances of secondary prominence have in common with examples like (24) or (25b). A good candidate is the phrase "mein afrikanischer Freund" *(my African friend)* in (26), found in our corpus of spontaneous monologues (see also Figure 5).

(26) [. . .] der junge Mann [. . .] Das Visum musste leider abgelehnt werden, weil Herr Nwahiri – so heißt [R-GIVEN mein [L-NEW afrikanischer [L-NEW Freund]]] – . . .
'[. . .] the young man [. . .] The visa unfortunately had to be dismissed because Mr. Nwahiri – that's the name of my African friend – . . . '

Figure 5: Realisation of an epithet (R-GIVEN, L-NEW) – "mein afrikanischer Freund"

This expression is an example of what is called an *epithet* (Clark, 1977; Schlenker, 2005). Such expressions can usually be characterized as coreferential expressions (R-GIVEN) which consist of new lexical material (L-NEW). (In this case "my African friend" refers back to "the young man"). They are not identical with cases of second occurrence focus (which, as we said, are defined as combinations of given or predictable and contrastive features) but exhibit a similar combination of boosting and inhibiting factors. Epithets typically cannot be produced with a nuclear accent because this would block the interpretation that the intended referent has been mentioned before, but they may receive a secondary prominence (cf. Figure 5).

Finally, we tentatively assume that the realisation of the secondary accent in example 24, can be described by assuming a joint effect of the ALT-FS feature and the referential givenness of the *fact* to which the *that*-clause refers. But this surely is worth of closer examination.

6 Summary

We have presented an annotation system for information structure which combines the advantages of a detailed classification of information status with the categorial freedom necessary to determine the givenness, accessibility or novelty of all parts of a clause and, therefore, focus-background information.

An important improvement is the differentiation between lexical relations like synonymy and hyponymy which hold between lexemes or set-denoting categories, and anaphora-related notions such as coreference or bridging which target referential expressions. Rather than saying that an expression is "given" or "new" we are now able to express that, for instance, a given individual is referred to by means of new lexical material. We also support the use of a further information structural level, which we call *elicitated alternatives* and which captures contrastive and other alternative-related properties of focus that do not belong to the domain of information status.

We have applied the annotation system to experimental and corpus data, as well as to theoretical examples that are taken from the literature on second occurrence focus. We also have sketched in what manner the detailed annotations which our system allows can be used for investigating phenomena which are prosodically marked by secondary accents.

In general, the labeling scheme serves to facilitate empirical investigations of subtle information structural and prosodic phenomena whose details by and large evade people's introspective abilities.

References

Nicholas Asher and Alex Lascarides. Bridging. *Journal of Semantics*, 15:83–113, 1998.

Stefan Baumann and Arndt Riester. Lexical and Referential Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta and P. Prieto, editors, *Prosody and Meaning*, Trends in Linguistics. Mouton de Gruyter, Berlin, subm.

Stefan Baumann, Doris Mücke, and Johannes Becker. Expression of Second Occurrence Focus in German. *Linguistische Berichte*, (221):61–78, 2010.

David Beaver and Brady Clark. *Sense and Sensitivity. How Focus Determines Meaning*. Wiley & Sons, Chichester, 2008.

David Beaver and Dan Velleman. The Communicative Significance of Primary and Secondary Accents. submitted to *Lingua*, subm.

David Beaver et al. When Semantics Meets Phonetics: Acoustical Studies of Second Occurrence Focus. *Language*, 83(2), 2007.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó. SALTO: A Versatile Multi-Level Annotation Tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

Daniel Büring. Been There, Marked That – A Theory of Second Occurrence Focus. in a volume edited by Makoto Kanazawa and Christopher Tancredi, to appear.

Steve Cassidy and Jonathan Harrington. Multi-Level Annotation in the EMU Speech Database Management System. *Speech Communication*, 1-2(33):61–78, 2001.

Wallace L. Chafe. *Discourse, Consciousness, and Time*. University of Chicago Press, 1994.

Herbert H. Clark. Bridging. In P. Johnson-Laird and P. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 169–174. Cambridge University Press, 1977.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 1(20):37–46, 1960.

Caroline Féry, Gisbert Fanselow, and Manfred Krifka, editors. *The Notions of Information Structure*, volume 6 of *Interdisciplinary Studies on Information Structure*. Universitätsverlag Potsdam, 2007.

Michael Götze, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas adn Ruben Stoel, and Thomas Weskott. Information structure. In Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors, *Information Structure in Crosslinguistic Corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure*, number 7 in Working Papers of the CRC 632, Interdisciplinary Studies on Information Structure (ISIS), pages 147–187. 2007.

Martine Grice, Stefan Baumann, and Ralf Benzmüller. German Intonation in Autosegmental-Metrical Phonology. In Sun-Ah Jun, editor, *Prosodic Typology. The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, 2005.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht, 1998.

Michael Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

Michael A. K. Halliday and C. Matthiessen. *An Introduction to Functional Grammar*. Edward Arnold, London, 2004.

Shinshiro Ishihara and Caroline Féry. Phonetic Correlates of Second Occurrence Focus. In *Proceedings of the 36th Meeting of the North Eastern Linguistics Society*, 2006.

Manfred Krifka. Basic Notions of Information Structure. In Féry et al. (2007), pages 13–56.

Knud Lambrecht. *Information Structure and Sentence Form*. Cambridge University Press, 1994.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. An Annotation Scheme for Information Status in Dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon, 2004.

Patrizia Paggio. Annotating Information Structure in a Corpus of Spoken Danish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1606–1609, Genoa, Italy, 2006.

Barbara H. Partee. Focus, Quantification and Semantics-Pragmatics Issues. In P. Bosch and R. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 213–231. Cambridge University Press, 1999.

Ellen F. Prince. Toward a Taxonomy of Given-New Information. In P. Cole, editor, *Radical Pragmatics*, pages 233–255. Academic Press, New York, 1981.

Ellen F. Prince. The ZPG Letter: Subjects, Definiteness and Information Status. In W. C. Mann and S. A. Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. Benjamins, Amsterdam, 1992.

Arndt Riester, David Lorenz, and Nina Seemann. A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722, Valletta, Malta, 2010.

Julia Ritz, Stefanie Dipper, and Michael Götze. Annotation of Information Structure: An Evaluation Across Different Types of Texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2137–2142, Marrakech, Morocco, 2008.

Craige Roberts. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *OSU Working Papers in Linguistics*, 49, 1996.

Christine Röhr and Stefan Baumann. Prosodic Marking of Information Status in German. In *Proceedings of Speech Prosody*, Chicago, 2010.

Christian Rohrer and Martin Forst. Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genova, 2006.

Mats Rooth. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1):75–116, 1992.

Mats Rooth. Second Occurrence Focus and Relativized Stress F. In C. Féry and M. Zimmermann, editors, *Information Structure: Theoretical, Typological, and Experimental Approaches*. Oxford University Press, 2010.

Phillippe Schlenker. Minimize Restrictors! (Notes on Definite Descriptions, Condition C and Epithets). In E. Maier, C. Bary, and J. Huitink, editors, *Proceedings of Sinn und Bedeutung IX*, pages 385–416, Nijmegen, 2005.

Roger Schwarzschild. GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177, 1999.

Elisabeth Selkirk. Contrastive Focus, Givenness and the Unmarked Status of 'Discourse-New'. In Féry et al. (2007), pages 125–145.

Extending Fine-Grained Semantic Relation Classification to Presupposition Relations between Verbs

Galina Tremper and Anette Frank
Department of Computational Linguistics
Heidelberg University, Germany

Abstract

In contrast to typical semantic relations between verbs, such as antonymy, synonymy or hyponymy, presupposition is a lexical relation that is not very well covered in existing lexical resources. It is also understudied in the field of corpus-based methods of learning semantic relations. But presupposition is very important for the quality of automatic semantic and discourse analysis tasks. In this paper we present a corpus-based method for learning presupposition relations between verbs, embedded in a discriminative classification approach for fine-grained semantic relations. The focus of the present paper is to discuss methodological aspects of our approach including the choice of resources and data sets, the selection of features for classification, and design decisions regarding the annotation of fine-grained semantic relations between verbs.

1 Introduction

Determining lexical-semantic and discourse-level information is crucial in event-based semantic processing tasks. This is not trivial, because significant portions of content conveyed in a discourse may not be overtly realized. Consider the examples (1) and (2), where (1) bears a presupposition that is overtly expressed in (2):

(1) *Spain won the finals of the 2010 World Cup.*

(2) *Spain played the finals of the 2010 World Cup.*

The presupposition expressed in (2) is implicitly encoded in (1), through lexical knowledge about the verb *win*, and is thus automatically understood by humans who interpret (1), given their linguistic knowledge about the verbs *win* and *play*.

One reason for addressing presupposition detection as a discriminative classification task is that *presupposition* needs to be carefully distinguished from other lexical relations, in particular *entailment* - as the two relations are closely related, but crucially differ in specific aspects. Consider the sentence pair (3) and (4).

(3) *The president John F. Kennedy was assassinated.*

(4) *The president John F. Kennedy died.*

Sentence (3) logically entails (4). But how does this differ from the presuppositional relation between (1) and (2)?

The differences between *presupposition* and *entailment* can be studied using special presupposition tests (Levinson, 1983). The most compelling one, which we will use throughout, is the negation test. It shows that the presupposition relation is preserved under negation, while entailment is not. Applied to (1) and (3), we note that (5), the negation of (1), still implies (2), while (6), the negation of (3), does not imply (4):

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 129-144, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

(5) *Spain didn't win the finals of the 2010 World Cup.*

(6) *The president John F. Kennedy was not assassinated.*

This can be taken as evidence that *win* presupposes *play*, while *assassinate* logically entails *die*. Thus, the negation test not only helps us to distinguish these closely related verb relations, it also points to the crucially distinct logical behaviour of these relations in deriving implicit meaning from discourse, which is the main motivation underlying our work.

Similar to entailment, presuppositional relations between verbs are essentially grounded in world knowledge. At the same time, they are crucial for the computation of discourse meaning and inference, and thus, need to be captured in large-scale lexicons, along with more structural, taxonomic semantic relations, such as *antonymy*, *synonymy*, or *hyponymy*. The latter are the primary relations that make up the WordNet database (Fellbaum, 1998). By contrast, *entailment*, *presupposition* and other more fine-grained relations are not covered in sufficient detail. Chklovski and Pantel (2004) were first to attempt the automatic classification of fine-grained verb semantic relations, such as *similarity*, *strength*[1], *antonymy*, *enablement*[2] and *happens-before* in VerbOcean. In the present paper we aim to extend the classification of semantic relations between verbs to *presupposition*. To our knowledge, it has not been attempted before. We will address this task in a discriminative classification task – by distinguishing presupposition from other semantic relations, in particular *entailment*, *temporal inclusion* and *antonymy*.

Our overall aim is to capture implicit lexical meanings conveyed by verbs, and to use this knowledge by making it explicit for improved discourse interpretation. This overall aim can be divided into two tasks:

Detecting and discriminating fine-grained semantic relations: We first detect and distinguish between fine-grained semantic relations including presupposition, at the type level, to encode this lexical knowledge in lexical semantic resources.

Deriving implicit meaning from text: In a second step, we will apply this knowledge for the interpretation of discourse, at the token level, in order to enrich the overtly expressed content with implied, implicit knowledge, conveyed by presupposition, entailment, or other lexical semantic relations. This kind of information can contribute to improving the quality of automatic semantic and discourse processing tasks, such as information extraction, text summarizarion, question-answering and full-fledged textual inferencing or natural language understanding tasks.

In the present paper we concentrate on the first task. We present a corpus-based method for learning semantic relations between verbs with a special focus on presupposition. The structure of the paper is as follows: Section 2 reviews related work.

[1] *strength*: V_1 are V_2 similar, but V_1 denotes a more intense action (Chklovski and Pantel, 2004)

[2] *enablement*: V_1 makes V_2 possible (Barker and Szpakowicz, 1995)

Section 3 studies the space and discriminative properties of fine-grained semantic relations, and introduces the basic method and selected features for classification and annotation strategies. In Section 4, we report our classification experiments. We introduce the resources used and discuss different annotation strategies. We present two corresponding classification experiments and the results we obtain. Section 5 offers a detailed error analysis regarding the used resources, features and annotation design schemes. Finally, we summarise and present objectives for future work in Section 6.

2 Related Work

Significant progress has been made during the last decade in the automatic acquisition of semantic relations between verbs using corpus-based methods. Lin and Pantel (2001) proposed a distributional method for extracting highly associated verbs. This method retrieves verb pairs which are linked by a semantic relation, but does not identify the type of these semantic relations. Their work was used as a starting point to automatically classify fine-grained semantic relations in other projects, such as VerbOcean (Chklovski and Pantel, 2004). Chklovski and Pantel (2004) used a semi-automatic pattern-based approach for extracting fine-grained semantic relations between verbs, including *similarity*, *strength*, *antonymy*, *enablement* and *happens-before*.

In a related strand of work, many projects tried to generate textual entailment rules (e.g. Pekar (2008), Ben Aharon et al. (2010)), however, they do not subclassify the extracted entailment pairs in *presupposition*, *entailment*, *cause* or other classes. Berant et al. (2010) try to improve on learning isolated textual entailment rules.

Only little work is devoted to the computational treatment of presupposition. Bos (2003) adopted the algorithm of van der Sandt (1992) for presupposition resolution. His approach is embedded in the framework of DRT (Kamp and Reyle, 1993). It requires heavy preprocessing and a lexical repository of presuppositional relations. Clausen and Manning (2009) compute presuppositions in a shallow inference framework called "natural logic". Their account is restricted to computing factivity presuppositions of sentence embedding verbs. In the field of corpus-based learning of semantic relations, the automatic acquisition of presupposition relations remains understudied.

3 A Corpus-based Method for Learning Semantic Relations

We present a corpus-based method for learning semantic relations between verbs including the presupposition relation. We subclassify verb pairs into five classes of relations: *presupposition*, *entailment*, *temporal inclusion*, *antonymy* and *other/unrelated*. The verbs in the last class stand in no or some semantic relation not considered here. In a preliminary step, we also considered the *synonymy* relation.

For classification, we start with a small number of seed verb pairs selected manually for each semantic relation and used to build a labeled corpus for training of binary

feature-based classifiers, one for each semantic relation. These classifiers are applied to a large set of unlabeled verb pairs. The candidate verb pairs are selected from a set of semantically related verbs according to the DIRT collection (Lin and Pantel, 2001). For the chosen candidate verb pairs, we extract corpus samples for feature-based classification in which both verbs co-occur, using the ukWaC corpus (Baroni et al., 2009). At this step we excluded the synonymy relation, as synonymous verbs usually do not occur contiguously in a single sentence.[3] For the remaining five semantic relations, we independently train five binary classifiers, using the J48 decision tree algorithm (Witten and Frank, 2005). Each of the five classifiers is applied to each sentence from the unlabeled corpus. The predictions of the classifiers are combined using ensemble learning techniques to determine the most confident classification.

3.1 Properties of Semantic Relations between Verbs and Feature Set

In order to establish an effective feature set for the classification we analyzed the properties of the relations between the verbs we aim to distinguish: *presupposition*, *entailment*, *temporal inclusion*, *antonymy* and *synonymy*[4]. We observe that the paradigmatic lexical semantic relations like *antonymy*, *synonymy* and *temporal inclusion* typically do not involve a temporal order. In contrast, *presupposition* relations between verbs involve a temporal sequence. The event that is presupposed tends to precede the event that triggers the presupposition. The verbs which stand in an *entailment* relation may or may not involve a temporal order; in case of temporal sequence the overtly realized verb can precede or succeed the entailed verb.

Another important aspect is the behaviour of the different semantic relations under negation. Some semantic relations (e.g. *presupposition* and *temporal inclusion*) are preserved under negation. In this way they can be distinguished from other semantic relations (e.g. *entailment* or *synonymy*) which do not persist under negation.

		Behaviour under Negation			
		$V_1 \rightarrow V_2$	$\neg V_1 \rightarrow V_2$	$V_1 \rightarrow \neg V_2$	$\neg V_1 \rightarrow \neg V_2$
Temporal Sequence	V_1 precedes V_2	E			E
	V_1 succeeds V_2	P E	P		P E
	No temporal sequence	E T S	 T A	 A	E T S

Table 1: Properties of the Semantic Relations:
P(resupposition), E(ntailment), T(emporal Inclusion), A(ntonymy), S(ynonymy)

[3] An analysis of sentences in which synonymy were found to co-occur shows that the verbs appear only accidentally within a single sentence, and should therefore be classified as unrelated. We therefore eliminated synonymy from the set of target relations.

[4] While we will classify synonyms as unrelated in our experiments, for completeness we do include synonymy in this discriminative analysis.

The distinguishing temporal and negation properties that cross-classify these semantic relations are schematically represented in Table 1[5]. As shown in Table 1, it is possible to distinguish the targeted semantic relations on the basis of these properties:

(i) *Presupposition* and *entailment* (whether or not temporally related) are distinguished on the basis of persistence under negation, which holds for *presupposition* only. The same pattern holds for *temporal inclusion* vs. *entailment*.
(ii) *Temporal inclusion* and *presupposition* behave alike regarding negation properties, but can be distinguished in terms of temporal sequencing properties.
(iii) *Synonymy* and *entailment* are difficult to distinguish in cases where *entailment* does not involve temporal sequence. However, since we exclude *synonymy* and range it under the class *unrelated*, this does not cause a problem.
(iv) *Antonymy* behaves clearly different from *entailment* and *presupposition* wrt. both properties, and from *temporal inclusion*, regarding negation properties.
(v) For completeness, *antonymy* and *synonymy* are opposites to each other wrt. negation properties, if we considered *synonymy* as a target relation.

Thus, the properties pointed out above could be used to distinguish the four target semantic classes. These four classes, in turn, need to be distinguished from the fifth class of unrelated verb pairs - which will include synonymous verbs, in case they (accidentally) co-occur. That is, we will need to model contextual relatedness features, to distinguish between the target relation classes and the class of unrelated verb pairs, and accidentally co-occuring verb pairs. For this purpose we will propose rather abstract contextual boundedness features that are able to characterize a broad variety of constructions that may be indicative for (any of) the targeted relation classes. We will refer to these features as "contiguity features".

3.2 Features for Classification

Temporal Sequence. To detect the distinct temporal relations between verbs we made use of features similar to the feature set used by Chambers et al. (2007):

1. Verb form (tense, aspect, modality, voice, negation, etc.).
2. PoS contexts (two words preceding and two words following each verb).

Further features we used for determining temporal relations are:

3. Coordinating/subordinating conjunctions.
4. Adverbial adjuncts.

Negation. Our analysis of the properties of the semantic relations shows that negation is crucial for distinguishing our target relations. Currently, we use as a trigger for

[5]Example of using the table: V_1 is a placeholder for the trigger verb and V_2 — for the inferred verb. For the *presupposition* verb pair *(win,play)*, the event of winning sth (V_1) typically temporally succeeds the event of playing something (V_2), therefore we concentrate on the second row. The event of winning something implies the event of playing something ($V_1 \rightarrow V_2$). The event of not winning something could be interpreted in two ways: constancy under negation — not winning although playing ($\neg V_1 \rightarrow V_2$) or cancellation — not winning because of not playing ($\neg V_1 \rightarrow \neg V_2$).

the negation feature the presence or absence of the negative particle *not/n't* (as part of the verb complex). In future work we plan to integrate further negation properties such as negation adverbs or suffixes.

Contiguity. One important task in the subclassification of verb relations is to decide whether or not two verbs stand in one of the targeted meaning relations in a given context. We observed that besides the distance between the verbs, the co-referential binding of the verb arguments can be very useful in determining contextual contiguity of verb pairs in specific contexts. Finally, in case of ambiguous verb readings, subcategorisation frames help to restrict a given verb relation to specific verb meanings. The following features were investigated for this purpose:

1. The distance between two analyzed verbs and the order of their appearance.
2. The number of main verbs occurring between two analyzed verbs.
3. The length of the path of grammatical functions relating the two verbs.
4. Co-reference relation holding between the subjects and objects of the verbs (both verbs have the same subject/object, the subject of one verb corresponds to the object of the second or there is no relation between them).
5. Subcategorization frames for two analyzed verbs.

4 Experiments and Results

4.1 Resources

In our experiments, we made use of the following resources.

1. *ukWaC* is the English part of the WaCKy corpora (Baroni et al., 2009). The corpus was constructed by crawling the .uk Internet domain and contains more than 2 billion tokens. Currently, it is the largest freely available resource for English that includes PoS and lemmatisation information. We use this corpus for extracting the training and test data sets, because it is large enough for obtaining high precision corpus data using statistical methods. ukWaC is certainly smaller than the entire English Web, but given that it is enriched with PoS and lemma annotations, multiple Internet queries can be replaced by a single one that applies to the pre-analysed ukWaC corpus. In our experiments we used the first three parts of the ukWaC corpus which contain about 280 million sentences.
2. Taking into account all possible combinations of verbs acquired from *ukWaC* yields an extremely large set of candidate pairs for classification, and the amount of unrelated verbs pairs would be huge. Instead, we used the DIRT-Collection to select pairs of highly associated verbs as candidates for classification. The *DIRT-Collection* (Lin and Pantel, 2001) is the output of the paraphrasing algorithm called DIRT applied on 1 GB of newspaper text from the TREC collection. It consists of pairs of verbs that have been determined to stand in a semantic relation using corpus-based association measures. DIRT contains 5,604 verb types and 808,764

verb pair types. We filtered the verb pairs extracted from DIRT using a threshold applied on the verb pair frequencies of appearance[6] and applied the PMI test with threshold 2.0. This reduces the number of candidate verb pairs to 199,393.

4.2 Annotation strategies for establishing a Gold Standard

Annotating semantic relations, especially implicit relations like *presupposition* and *entailment* is a difficult task because of the subtlety of the tests and the involved decisions. In order to obtain reliable annotations it is important to define the task in an easy and accessible way and to give clear instructions to the annotators.

We decided to formulate two annotation tasks: one on the level of verb pairs given as types out of context (type-based annotation) and another on the level of verb pairs in context (token-based annotation) and to examine to what degree the results obtained from the two annotation setups correlate.

4.2.1 Gold Standard 1 (GS1): Type-based annotation

The complete set of verb pair candidates (about 200,000 verb pairs) is impossible to annotate manually, therefore we randomly selected a small sample of 100 verb pairs. In order not to influence the judges' decisions, we eliminated the system annotations. Since some verbs can have more than one meaning and consequently verbs in a verb pair can stand in more than one semantic relation, the judges were allowed to assign more than one relation to each verb pair.

To support the annotators in their decisions, we provided them with a couple of inference patterns and examples for each semantic relation. This is shown in Table 2.

Sem. Relation	Pattern	Example	*Substitution in pattern*
Presupposition	V_1 presupposes V_2, *not* V_1 presupposes V_2	*win - play*	*winning* presupposes *playing* *not winning* presupposes *playing*
Entailment	V_1 implies V_2, *not* V_1 doesn't imply V_2	*kill - die*	*killing* implies *dying* *not killing* doesn't imply *dying*
Temporal Inclusion	V_1 happens during V_2 or V_1 is a special form of V_2	*snore - sleep* *mutter - talk*	*snoring* happens during *sleeping* *muttering* is a special form of *talking*
Antonymy	either V_1 or V_2, V_1 is the opposite of V_2	*go - stay*	either *going* or *staying* *going* is the opposite of *staying*
Other/unrelated	none of the above	*jump - sing*	

Table 2: Semantic Relations and Inference Patterns for Annotation

The inter-annotator agreement for this task was 63% corresponding to a Kappa value of 0.47. This can be taken as an indication for a high difficulty of semantic relation

[6]The verb pair frequencies were calculated only for the first three parts of the ukWaC corpus.

annotation when performed out of context.

4.2.2 Gold Standard 2 (GS2): Token-based annotation

Since type-based annotation turned out to be very difficult, we decided to simplify the task by providing the annotators with verb pairs in their original context. For this token-based annotation we took the same 100 verb pairs and randomly selected 5 to 10 contexts for each of them (the total number of all contexts was equal to 877). Similar to the type-based annotation task we eliminated all system labels. In contrast to type-based annotation, we only accepted a single relation label for a given verb pair.

The inter-annotator agreement for this task was 77.4%, which corresponds to a Kappa value of 0.44. Error analysis showed that the most important problems are not due to semantic relations which are difficult to distinguish (e.g. *presupposition* and *entailment*), but rather in determining whether or not there is a specific semantic relation between two verbs in a given context.

We examined the correlation between the type- and token-based Gold Standards by comparing the annotations of a single judge for both annotation tasks. For 62% of verb pair types we observe an overlap of labels, 28% of the verb pair types were assigned labels on the basis of the annotations in context which were not present on the type level without context, or the type level label was not assigned in context, because of the small amount of contexts for a verb pair. For 10% of verb pair types we found conflicting annotations (for example, *presupposition* and *entailment*). Thus, for the most part (62%) the type-based annotation conforms with the ground truth obtained from token-based annotation. An additional 28% of verb pairs can be considered to be potentially correct. The divergences for these verb pairs could be explained by the random procedure of the context extraction which does not always return appropriate contexts. They can also be explained by the difficulty for the annotator to consider all possible verb meanings for highly ambiguous verbs in type-based annotation.

4.2.3 Gold Standard 3 (GS3): Type-based annotation deduced from GS2

Since our ultimate goal is to detect and distinguish fine-grained semantic relations at the type level, we used the token-based annotations to deduce type-based annotations. For GS1 we accepted multiple relation labels. Therefore, for constructing GS3, for each verb pair type we selected up to three most probable annotations (most frequent annotations from the token-based annotations of GS2). An exception was made for the *other/unrelated* class: only the verb pairs annotated unambiguously in all cases with the *other/unrelated* label were considered to belong to this class.

The distribution of semantic relations in all three Gold Standards is given in Table 3.

Semantic Relation	Frequency in GS1	Frequency in GS2	Frequency in GS3
Presupposition	18	70	24
Entailment	8	44	8
Temporal Inclusion	19	26	12
Antonymy	12	44	10
Other/unrelated	43	693	46

Table 3: Distribution of Semantic Relations in Gold Standards (GS)

The distribution of relation types in GS1 and GS3 is very close. Because GS3 was derived from GS2 by selecting up to three most probable annotations, the overlap between them is identical to the overlap between GS1 and GS2 discussed above (62%)[7]. A confusion analysis shows that the set of verb pairs labeled as *entailment* remains stable (*entailment* and *presupposition* are confounded in only two cases). Annotation in context reduces the number of *temporal inclusion* and *antonymy* relations that were annotated out of context. On other hand, we observe a tendency to annotate more verb pairs with the *presupposition* relation.

4.3 Best Features for Classification

Our final classification is based on five binary classifiers, one for each semantic relation. We analyzed which of the features from the feature set (see Section 3.2) are the most effective for determining each semantic relation. We also compared the best features for binary classifiers with the best features for single multi-class classification. The best features were determined on the basis of the manually annotated training set using Gain Ratio coefficient. The top five best performing features for each individual semantic relation and for the full set of relations are presented in Table 4.

Table 4 shows that conjunctions between verbs are important for all semantic relations. For determining presupposition, the verb that triggers the presupposition (V_1) seems to be more important than the presupposed verb (V_2). By contrast, for determining the entailment relation, the verb which is the logical consequence (V_2) seems to be more important than the verb which implies it (V_1). The selected features highlight the importance of coreference relations holding between arguments, as well as the subcategorization frame information for detecting a specific semantic relation between verbs. They characterize in particular the unrelated class, and *antonymy*, as contextually unrelated verbs[8]. Negation was not selected as a strong feature, although it prominently figures in our analytical cross-classification scheme. This may be due to sparseness,

[7]For computing overlap we consider all relations annotated per type.

[8]This suggests exploring a two stage classification that in a first step distinguishes unrelated verbs from related ones, and subsequently classifies the remaining fine-grained semantic relations.

Semantic Relation	Top-5 Best Features
Presupposition	Order, Conj, AdvAdj of V_1, Mod of V_1, SubCat of V_1
Entailment	Order, Conj, AdvAdj of V_2, Mod of V_2, Asp of V_2
Temporal Inclusion	Conj, AdvAdj of V_1, SubCat of V_1, SubCat of V_2, Dist
Antonymy	Conj, AdvAdj of V_2, SubjObj, NumVerb, Dist
Other/no	Conj, SubjObj, SubCat of V_1, SubCat of V_2, GF-Length
All	Order, Conj, AdvAdj of V_1, SubCat of V_1, SubjObj

Table 4: Top-5 Best Features
V_1, V_2 – placeholders for verbs in the verb pair, Order – Order of appearence, Conj – Conjunction, AdvAdj – Adverbial adjunct, Mod – Modality, Asp – Aspect, Dist – Distance between verbs, GF-Length – length of GF-path between verbs, NumVerb – number of intervening main verbs, SubCat – Subcat frame, SubObj – Coreference between Subject/Object

given the restricted feature set currently used for characterizing negation properties.

4.4 Classification

Starting with a small number of seed verb pairs (3 to 6) (see beginning of Section 3), we build a training corpus consisting of three parts: a manually annotated training set (5,032 sentences) collected from the ukWaC for the seed verb pairs, a heuristically annotated extended training set (9,918 sentences)[9] and heuristically annotated synonymous verb pairs in context (757 sentences)[10]. The set of unlabeled verb pairs in context is built from the filtered set of related verb pairs from DIRT (see Section 4.1), and includes about 4,500,000 sentences. For the classification we use the outputs of five binary J48-classifiers independently applied on the same set of unlabeled data.[11]

4.5 Experiments and Results

We performed two experiments for the classification of verb pairs. In the first experiment we classified each candidate verb pair in context (token-based classification) and evaluated the results against GS2. In the second experiment we classified all candidate verb pairs at the type level, by aggregation from instance-level classifications in context (type-based classification) and evaluated the results against GS1 and GS3.

[9] Heuristic annotation was performed using a manually compiled small stoplist of patterns meant to eliminate wrong instances (see Tremper (2010) for details). In future work we will explore the use of classifiers trained on shallow features (Banko et al., 2007).

[10] The synonyms were obtained from WordNet (Fellbaum, 1998).

[11] We also experimented with classification using a multiclass J48-classifier. Due to the lower results on a small subset of the manually annotated training corpus, we didn't evaluate this classifier on the unlabeled data set.

4.5.1 Experiment 1

To perform **token-based classification** we determined the most confident classification for each instance of the unlabeled verb pair in context using a voting architecture. We compared the classifications of all five binary classifiers and selected the classification with the highest confidence.[12]

We evaluated the results against the token-based Gold Standard 2 (see Section 4.2.2). We computed precision, recall and f-score. As baseline we took the distribution found in the manually labeled gold standard as the underlying verb relation distribution. The results for each semantic relation are shown in Table 5.

Semantic relation	Precision	Recall	F-Score	Baseline
Presupposition	23%	27%	25%	8%
Entailment	18%	25%	21%	5%
Temp. Inclusion	10%	12%	11%	3%
Antonymy	42%	68%	52%	5%
Other/Unrelated	73%	59%	65%	79%
Macro-Average	33%	36%	34%	
Micro-Average	59%	54%	56%	

Table 5: Evaluation of the Results of Experiment 1

Except for the unrelated class, the results are well above the baseline. The results show that typical and broad semantic relations such as *antonymy* perform better than *presupposition* and *entailment*. *Temporal inclusion* achieves the lowest results for token-based classification. Here, the error analysis shows that this relation was most often confounded with unrelated verb pairs. Some examples of the correct and wrong classifications are presented in the Table 6.

4.5.2 Experiment 2

To perform **type-based classification** we first performed token-based classification as described in Experiment 1. We combined the results obtained for individual instances to derive relation labels on the type level as follows. We eliminated semantic relation labels which were assigned to less than 10% of the instances of a given verb pair. Verb pairs which after this step were assigned more than three semantic relations are ignored (remain unclassified). Finally, verb pairs that were left with up to three semantic relations, each of which was assigned to at least 10% of the examples, were labeled with all of these semantic relations.

We evaluated the results against the type-based Gold Standard 1 (see Section 4.2.1) and Gold Standard 3 (see Section 4.2.3). Again we report precision, recall and f-score.

[12]Only the classifications with a confidence exceeding 0.75 were accepted for voting.

Sem. Relation	Verb pair	Correct classification	Wrong classification (System label)
Presupposition	*classify – identify*	It was noted that of the thirteen issues identified in the report eight were classified as high priority.	The meeting focussed on issues of identifying, classifying and marking up names in both corpora and analytical projects. (Temp. Inclusion)
Entailment	*click – send*	Clicking the Send feedback button will send any feedback you have entered.	You can send us your comments by simply clicking on this email. (None)
Temp. Inclusion	*reply – say*	Replying to the toast to the guests, Dr Julia King said how privileged the Faculty was to have two such active alumni associations.	18 out of the 20 Rehabilitation Officers who replied said that there is somewhere they can take clients for equipment demonstrations. (None)
Antonymy	*disconnect – connect*	A click should be heard every time the antenna wire is connected or disconnected.	This allows you to connect and disconnect easily , simply by clicking on the icon and selecting the relevant option. (None)

Table 6: Examples of the correct and wrong classifications in context

In contrast to token-based classification, we considered verb pairs to be correctly labeled if at least one tag was correct. The results are shown in Table 7.

Semantic relation	Gold Standard 1				Gold Standard 3			
	Prec.	Recall	F-Score	Baseline	Prec.	Recall	F-Score	Baseline
Presupposition	43%	33%	37%	18%	50%	29%	37%	24%
Entailment	36%	50%	42%	8%	36%	50%	42%	8%
Temp. Inclusion	50%	16%	24%	19%	33%	17%	22%	12%
Antonymy	75%	75%	75%	12%	58%	70%	63%	10%
Other/Unrelated	56%	74%	64%	43%	68%	85%	76%	46%
Macro-Average	33%	50%	40%		49%	50%	49%	
Micro-Average	53%	53%	53%		59%	59%	59%	

Table 7: Evaluation results for Experiment 2 (against Gold Standards 1 and 3)

The type-based classification clearly outperforms token-based classification. One of the reasons for the better performance of type-based classification is certainly that more examples are considered for assigning a relation, in which case voting plays a major role in eliminating unsecure decisions. By contrast, in token-based classification, each example is considered and labeled in isolation, including those with small confidence scores. The results for type-based classification are clearly exceeding the baseline for all relation types. Comparing evaluation against GS1 and GS3, the results for GS3 are higher, with a balanced macro-average in precision, recall and f-scores of about

50%, with clear improvement of precision for *presupposition*, a drop in performance for *antonymy*, and high performance gains for distinguishing the *unrelated* class.

5 Error Analysis and Discussion

5.1 Resources

Using the verb pairs extracted from the DIRT collection (see Section 4.1.), we extracted corpus samples from the ukWaC corpus (both for establishing labeled training and unlabeled test corpora). The PoS and lemma information encoded in ukWaC saves time needed to tag and lemmatise the corpus. But it also incurs errors that cause problems in the classification. An error analysis conducted on a small subset of the manually annotated training corpus shows that 10% of all errors are caused by erroneously annotating nouns or adjectives as verbs. This problem can be solved by using information from a deep parser to double check the PoS-tags provided by ukWaC.

5.2 Annotation

Comparing the results of the two annotation setups clearly shows that both are difficult, yet in different ways. Annotation on the type level is difficult because no indication is given about the intended meaning of the verbs. Hence the annotators need to consider all possible combinations of meanings for any pair of verbs. However, embedding the pairs in their original context doesn't make the decision much easier. This is because some sentences involve complex structure and interpretation difficulties, which require a lot of attention and time to annotate the individual examples.

To render the annotation task more reliable and less time-consuming, we need to develop an annotation strategy which includes the positive elements of both annotation strategies described above. One solution could be to present verb pairs with prototypical arguments instead of taking the concrete sentence as a disambiguating context. The argument abstractions could be represented using WordNet hypernyms.

Another strategy could be to use a question scenario to collect annotations. The idea is to guide the annotator to verifying the discriminative categorizing properties "temporal sequence" and "persistence under negation", using a "setting" and a follow-up question that is intended to elicit the critical/missing piece of information needed to classify the verb pair in question. Using the properties of the semantic relations displayed in Table 1 we established a set of questions that elicit only three possible answers (Yes/No/Maybe). The answers can be used to distinguish between the target semantic relations and thus to annotate the data. (7)-(9) list examples of such questions:

(7) ***X** found **Y**. Did **X** search **Y**?*[13]

The answer *yes* in (7) excludes the semantic relation *antonymy* for the pair *find* and *search* (as antonyms can't be valid at the same time).

[13] **X** and **Y** in the questions are placeholders for arguments which can be refined using prototypical nouns.

(8) ***X** didn't find **Y**. Did **X** search **Y**?*

The answer *maybe* in (8) indicates persistence under negation, and thus excludes the relation *entailment*.

(9) *Did **X** find **Y** after searching?*

The answer *yes* in (9) excludes the relation *temporal inclusion* between *find* and *search*. On the basis of these three answers we can annotate the verb pair with *presupposition*. By exploiting the properties of the target relations regarding temporal sequence and negation, as summarized in Table 1, we can distinguish each of the 5 target classes with maximally three questions per verb pair. The decision tree for distinguishing between semantic relations is presented in Figure 1.

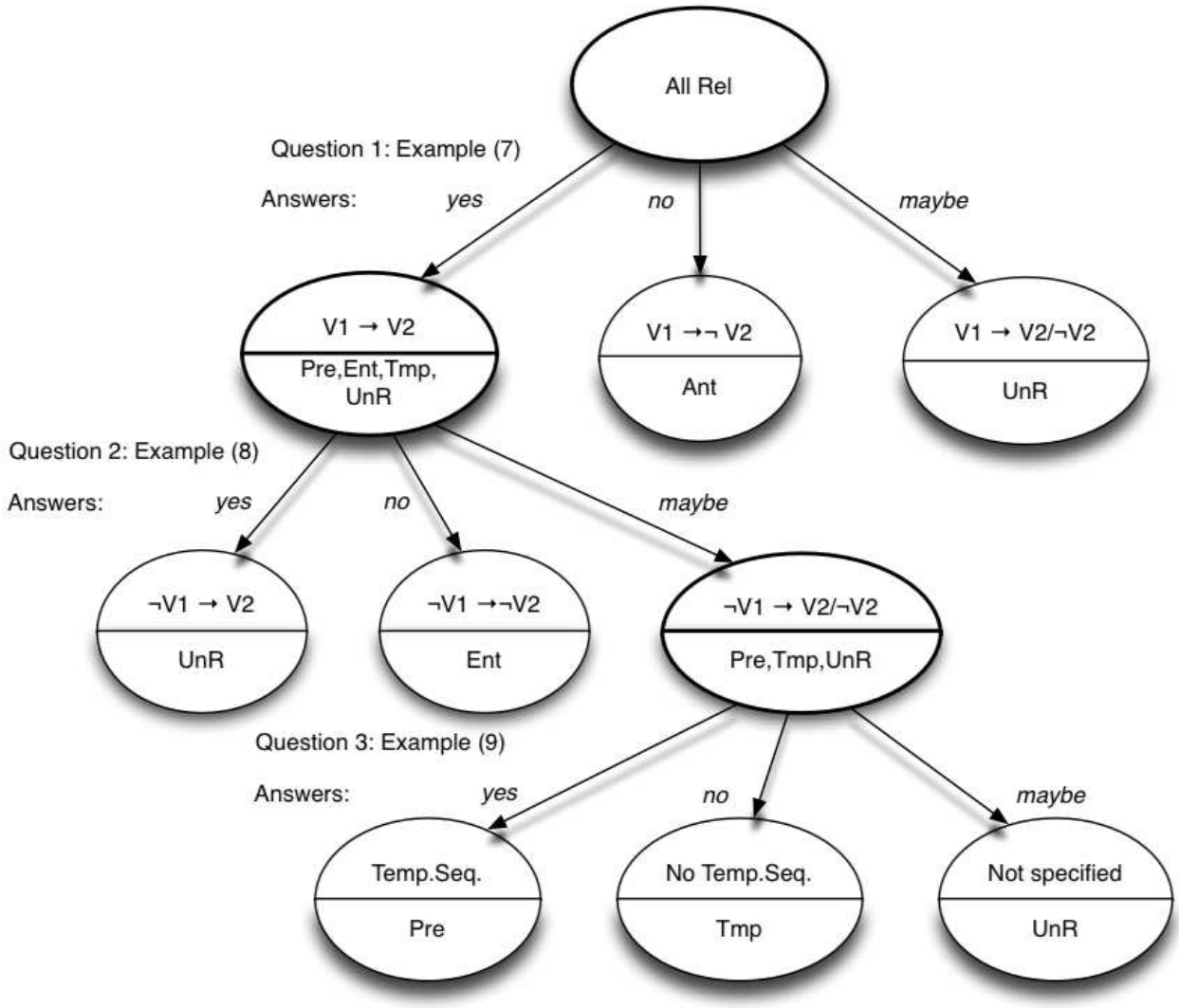


Figure 1: Decision Tree for distinguishing between semantic relations
Pre – presupposition, Ent – entailment, Tmp – temporal inclusion, Ant – antonymy, UnR – unrelated, Temp. Seq. – temporal sequence, V_1 & V_2 – placeholders for verbs

6 Conclusion and Future Work

In this paper we present first results in the corpus-based acquisition of presupposition relations between verbs, embedded in a discriminative classification approach for fine-grained semantic relation classification. We observe that presupposition is more difficult to determine than typical semantic relations like antonymy.

There are still many open issues left for future work. Coming up with solutions for word sense disambiguation and coreference resolution could help to eliminate the major source of observed errors. To improve the reliability of annotation and system performance, we plan to integrate information about predicate-argument structure using information extracted from FrameNet (Ruppenhofer et al., 2005) and VerbNet (Kipper, 2005) as well as prototypical argument head nouns encoding selectional preferences. We aim to improve classification performance by extending our feature set for characterizing negation properties. We also plan to evaluate the question-based annotation scenario proposed in Section 5.2. Given that it relieves the annotator from considering complex logical decisions, it could be appropriate for a crowd-sourcing annotation setup. We will also investigate a cascaded classification approach that follows the structure of the annotation decision tree.

The focus of the present paper was to describe in detail the underlying properties of the selected relations, our choice of resources and features for context-based classification, and to discuss design issues of the annotation task. Future work will establish an annotation and evaluation setup for the induction of implicit information in context, using the acquired semantic relations, in particular the presupposition relation pairs.

Acknowledgements

We would like to thank in particular our annotators: Carina Silberer, Eva Sourjikova and Matthias Hartung, and the anonymous reviewers for valuable feedback.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, Hyderabad, India, 2007.

Ken Barker and Stan Szpakowicz. Interactive semantic analysis of clause-level relationships. In *Proceedings of PCLING 95*, pages 22–30, Brisbane, Australia, 1995.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226, 2009.

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. Generating entailment rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246, Uppsala, Sweden, 2010.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. Global learning of focused entailment graphs. In *Proceedings of the ACL 2010 Conference*, pages 1220–1229, Uppsala, Sweden, 2010.

Johan Bos. Implementing the binding and accommodation theory for anaphora resolution and presupposition projection. *Computational Linguistics*, 29(2):179–210, 2003.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of the ACL-07 conference*, pages 174–176, Prague, Czech Republic, 2007.

Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, 2004.

David R. Clausen and Christopher D. Manning. Presupposed content and entailments in natural language inference. In *Proceedings of the 2009 Workshop on Applied Textual Inference, ACL-IJCNLP 2009*, pages 70–73, 2009.

Christian Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, first edition, 1998.

Hans Kamp and Uwe Reyle. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, 1993.

Karen Kipper. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

Stephen C. Levinson. *Pragmatics*. Cambridge: Cambridge University Press, 1983.

Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360, 2001.

Viktor Pekar. Discovery of event entailment knowledge from text corpora. *Computer Speech & Language*, 22(1):1–16, 2008.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. FrameNet II: Extended theory and practice. Technical report, ICSI, 2005. URL `http://framenet.icsi.berkeley.edu/book/book.pdf`.

Galina Tremper. Weakly supervised learning of presupposition relations between verbs. In *Proceedings of the ACL 2010, Student Research Workshop*, pages 97–102, Uppsala, Sweden, 2010.

Rob van der Sandt. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9:333–377, 1992.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Amsterdam, 2nd edition, 2005.

Towards Finer-Grained Tagging of Discourse Connectives

Yannick Versley
Universität Tübingen

Abstract
Many recent experiments in the automatic classification of discourse relations have limited themselves to a small set of coarse categories. While there are eminent reasons to do so – annotation for a small sets of categories can be created more reliably, and possibly also be defined in a more clear – cut way than finer distinctions – it is an interesting question whether the finer-grained distinctions present in some annotated corpora can be reconstructed reliably. The present paper investigates the feasibility of such fine-grained tagging of discourse relations using data from the Penn Discourse Treebank.

1 Introduction
In order to structure a discourse beyond the level of single clauses and the predicate-argument relations contained therein, speakers or writers implicitly or explicitly express relations between events, propositions, or speech acts expressed in different clauses – so called *discourse relations*. Often (but still in a minority of cases) such discourse relations are marked by *discourse connectives*, which signal the presence of a relation between arguments that can be determined either purely syntactically (in the case of coordinating or subordinating conjunctions) or anaphorically (e.g., in the case of discourse adverbials).

Early work on discourse parsing (Soricut and Marcu, 2003) has focused mostly on such overtly marked discourse relations – both because they are easier to detect in general and because the discourse connective itself considerably constrains the kinds of relations that can hold between its arguments. (Some connectives such as *although* always mark one kind of relation, whereas other connectives such as *since* or *and* are more ambiguous).

Later work such as Sporleder and Lascarides (2008); Pitler et al. (2009); Lin et al. (2009) focused on the sense disambiguation of implicit discourse relations, which is more sensitive to semantic information as the lack of an explicit connective yields a significantly higher ambiguity for the realized relation. However, classification accuracy on implicit discourse relations only reaches accuracies of 44.6% (Pitler et al., 2009, for the 4 upper-level categories of the Penn Discourse Treebank (PDTB) plus *EntRel* and *NoRel* for the non-presence of a discourse relation), or 40.2% (Lin et al., 2009, for the 16 mid-level PDTB categories), despite the fact that the (textual/discourse) units to be related are assumed as given. Therefore, methods using explicit cues are currently closer to being useful in actual applications.

Of the existing research on disambiguating the discourse relations signaled by connectives, Haddow (2005) and Miltsakaki et al. (2005) focus on a small number of ambiguous connectives, using a set of relations motivated by said set of connectives. In

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 145-155, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

contrast, Pitler and Nenkova (2009) consider the full range of discourse connectives present in the PDTB, which allows to gain a more comprehensive overall picture. Pitler and Nenkova report results only for the topmost level of the PDTB's relation inventory, which comprises four coarse relation types (*Comparison*, *Expansion*, *Contingency* and *Temporal*). As the PDTB (on the finer granularity levels of their label set) – as well as most other discourse corpora – includes finer distinctions, it may be of interest whether these finer distinctions can also be made automatically. Accurate classfication also at the lower level, using methods that assume only information that can be produced by automatic preprocessing, would clearly also be beneficial from an application perspective.

In this paper, we discuss the problems that are faced by discourse tagging in the finer distinctions of the relation taxonomy, and propose suitable methods for hierarchical classification that allow the prediction of finer classes while making use of the taxonomical information contained in the PDTB's hierarchical label set. We also discuss additional features that help in making these finer-grained distinctions more robustly.

2 Setting

2.1 Data: The Penn Discourse Treebank

The Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008) contains, for the text basis covered by the Wall Street Journal portion of the Penn Treebank, annotation of discourse relations marked by a connective (*Explicit*), those that are not marked by a connective (*AltLex* and *Implicit*), as well as annotations that do not signal a discourse relation (*EntRel* and *NoRel*). The present study focuses on the 15 366 *Explicit* relations found in the PDTB (or more specifically, its sections 2-22).

The four coarse relation types in the Penn Discourse Treebank (*Comparison*, *Expansion*, *Contingency* and *Temporal*) are further subdivided into sixteen second-level relations, among which ten occur more than 200 times within sections 2-22: Within the *Comparison* group, this includes the distinction between *Concession* and *Contrast*, within the *Contingency* group, the one between *Cause* and *Condition*, and the *Expansion* group includes, besides *Instantiation* as its predominant member, the *Alternative*, *Instantiation*, and *List* relations. Within the *Temporal* group, a further distinction is drawn between *Asynchronous* and *Synchrony* relations. (Among the infrequent second-level relations are the 'pragmatic' variants of concession, contrast, cause, and condition, all occurring 50 times or less, as well as *Expansion.Restatement*, which occurs 128 times).

The third level, finally, distinguishes multiple variants of *Contrast* and *Concession* based on the relation between the objects or propositions that are related, *Cause* contains the division between *Reason* and *Result* (corresponding to the causal ordering of the assigned arguments, which normally only varies with syntactic properties of

the connective), as well as various distinctions among *Condition* based on factuality; within the *Asynchronous* temporal relations, the third level distinguishes *Precedence* and *Succession*.

2.2 Features

Discourse classification is carried out as a supervised machine learning task, using features that summarize the linguistic properties of the discourse connective's context. Two of the used features are reimplementations of ones used by Pitler and Nenkova (2009): One is the string of the connective itself. In order to reduce the annotated spans in the *connector* slot of the annotation, which can include additional text to the connective itself (such as "*two minutes*" in "*two minutes before the train departed*"), occurrences ending with one of *after*, *before*, *when*, *until*, *since*, or *if* were shortened to that word whenever the span was longer. The case of all connectives was normalized to lower case.

The second group of features comprises Pitler and Nenkova's syntactic features: These include the labels of self, parent, left sibling and right sibling nodes (counting from the lowest node that covers all of the words annotated as connective span and that is not the only child of its parent), as well as additional features signaling the presence of a VP node or of a trace as a child of the right sibling.

In line with the observations by Pitler and Nenkova, we found that the ambiguity between *Temporal* relations and *Contingency* relations (specifically, *Condition*) was a major source of misclassifications. The main difference between *Temporal* and *Contingeny* relations in the explicit cases lie in the facticity of the connected events. Both Miltsakaki et al. (2005) and Haddow (2005) use additional features that pertain to tense and mood of the connected arguments, but presuppose the arguments as given.

To be able to use these features with automatic preprocessing and tell whether they are informative with respect to the distinction between *Temporal* and *Contingency* relations (as well as the accuracy of relations on the finer levels of the taxonomy), we automatically derive the argument nodes from the syntactic annotation of the treebank. While the PDTB annotation contains argument spans, methods for their automatic identification are not perfect – Elwell and Baldridge (2008) report accuracy scores of 82.0% and 93.7%, which means that using perfect information in the identification of discourse relations may create a distorted picture.

As a simple, high-precision mechanism to identify arguments, we implemented heuristics to derive the argument nodes using syntactic heuristics for different groups of connectives, in particular subordinating coordinators (*[S . . . [SBAR [IN after] she slept]]*), clausal PPs (*[S . . . [PP [IN after] [S sleeping]] . . .]*), sentence coordination (*[S [S he sleeps] [CC and] [S he snores]]*), w-adverbials (*[S . . . [SBAR [WHADVP when] he sleeps]]*), as well as fronted (preposition- or adverb-headed) adverbials, which have

one of their arguments (the ARG2 in PDTB parlance) in the current sentence whereas the other is linked anaphorically.

Based on the identified arguments, we extract the following indicators:

- the part-of-speech of the first non-modal verb in the sentence (descending from the argument clause node into further VP and S nodes to cover both nesting of VPs and coordinated sentences)
- the presence (and word form) of modals and negation in the clause
- a tuple of *(have-form, be-form, head-POS, modal present)* as proposed by Miltsakaki et al. (2005).

(In the result tables, the part-of-speech/presence of modals pair of features will be called *pos*, whereas the tuple describing auxiliaries, the POS of the lexical head, and the presence of modals will be simply called *verb*).

Verb tense and modals are relatively shallow correlates of more interesting properties such as facticity or veridicality (i.e., whether the speaker asserts the propositional content of that clause to be true), but they are easy to extract in a robust manner and useful as a first approximation to a more comprehensive approach such as those of Palmer et al. (2007) to classifying situation entities.

2.3 Hierarchical Classification

Considering that the Penn Discourse Treebank has a hierarchical label set, relevant generalizations may be found at multiple levels of the relation hierarchy. In the area of word sense disambiguation, Ciaramita et al. (2003) have shown that a classifier that uses a two-level hierarchy to generalize the word senses performs better than a state-of-the-art "flat" multiclass classifier.

For our version of the hierarchical classification, we start from a maximum entropy classifier (Berger et al., 1996), in contrast to Ciaramita et al., who use a Perceptron classifier.[1] In the standard formulation, maximum entropy learning minimizes the loss

$$\mathrm{Loss}(w) = \prod_{x,y} \log \frac{\mu(x,y)}{\sum_{y' \in Y} \mu(x,y')}$$

where the measure $\mu(x,y)$ is defined as

$$\mu(x,y) = \exp(\langle w, \phi(x,y) \rangle)$$

for a feature function ϕ that pairs all features extracted from x with the label for y.

In the hierarchical case, ϕ pairs the features extracted from x not only with the actual class label y, but also nodes higher up in the taxonomy - yielding, for example, not only

[1] A wide variety of learning algorithms can be used to learn linear multiclass classifiers such as those used by Ciaramita et al. and in this work, of which the standard techniques for maximum entropy estimation – optimizing a log-likelihood-based loss using quasi-Newton numerical optimization – are by far the most commonly used.

a weight for "x has an SBAR parent and y is *Contingency.Cause.Result*", but also for the more general "x has an SBAR parent and y is a descendent of *Contingency.Cause*". To improve the separability of the problem at hand, we consider combinations of up to two of the original features from x.

As the PDTB contains underspecified relations (e.g., just *Contingency*) in cases where annotators could not reach an agreement about the finer relation, such labels would occur as possible tags for relation instances, including those that are tagged with a finer label. To avoid the confusion that would arise from using the underspecified relations either as positive or negative example, we completely removed the less-specific relation from the learning instance if it was labeled with a more-specific relation.

To make use of the presence of multiple relation labels in the annotation of the Penn Discourse Treebank (for example, a given instance of a connective may receive *Temporal.Synchrony* as the primary classification and *Comparison.Contrast.Juxtaposition* as a secondary classification) we chose to optimize the (sum) probability that the model assigns to *all* of the correct labels:

$$\text{Loss}(w) = \prod_{x, Y_{good}} \frac{\sum_{y \in Y_{good}} \mu(x, y')}{\sum_{y' \in Y} \mu(x, y')}$$

Besides the flat multiclass classifier and the hierarchical classifier, we also implemented a method for *greedy* classification, where the top-level relation is determined and subsequent relations are determined by a specialized classifier that, for a given relation prefix, guesses the next element. For example, the topmost classifier would classify the relation as *Temporal*, then the second-level classifier for *Temporal* would determine that the relation is *Temporal.Asynchronous*, and the third-level classifier for *Temporal.Asynchronous* would choose *Temporal.Asynchronous.Precedence* as the finest-level relation.

3 Results

For the quantitative evaluation, we follow Pitler and Nenkova in treating the system classification as correct whenever it matches the label, or one of multiple assigned labels, from the manual annotation. To account for the underspecified relations in the PDTB, we also count the system response as correct when it is more specific than the gold-standard label (or one of the gold-standard labels) – for example, when the corpus annotation contains an underspecified *Comparison* annotation, but the system predicts *Comparison.Concession* or even *Comparison.Concession.Contraexpectation*, our evaluation would count this as correct.

Tables 1, 2, and 3 show the results for using different classification methods. Except for the 'greedy' classifier on the finer relations using the approximate tense/mood features, we see only very small differences on the order of 0.1-0.2%, with the greedy classifier performing slightly better on the finer relation levels.

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.946	0.954**	0.954**	0.954**	0.953**	0.952**	0.952**
d=2		0.840	0.839		0.847*	0.847**		0.845	0.845
d=3			0.790			0.796*			0.798**

Table 1: Flat classification

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.945	0.954**	0.953**	0.954**	0.953**	0.952**	0.952**
d=2		0.840	0.839		0.847**	0.847*		0.845	0.845*
d=3			0.790			0.796*			0.798**

Table 2: Hierarchical classification

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.946	0.955**	0.954**	0.955**	0.953**	0.953**	0.953**
d=2		0.840	0.840		0.847*	0.847*		0.845	0.845
d=3			0.792			0.798*			0.800*

Table 3: Greedy classification

Differences to connective-only version: significant at $p < 0.01$ (*) / significant at $p < 0.001$ (**)

Relation	N	Prec	Recl	F
Comparison	**4566**	*0.960*	*0.968*	*0.964*
Comparison.Contrast	3102	0.771	0.898	0.829
Comparison.Concession	1080	0.549	0.309	0.396
Contingency	**2634**	*0.970*	*0.873*	*0.919*
Contingency.Cause	1456	0.982	0.868	0.921
Contingency.Condition	1123	0.919	0.883	0.901
Expansion	**5206**	*0.979*	*0.960*	*0.969*
Expansion.Conjunction	4293	0.920	0.955	0.920
Expansion.Alternative	300	0.926	0.914	0.920
Expansion.Instantiation	245	0.992	0.963	0.977
Expansion.List	205	0.000	0.000	0.000
Temporal	**2961**	*0.882*	*0.966*	*0.923*
Temporal.Asynchronous	1712	0.938	0.869	0.902
Temporal.Synchrony	1244	0.691	0.937	0.795

Table 4: Results for the most frequent second-level relations (connective+syntax)

As can be seen in Table 4, the only problem at the coarsest level of relations is a misclassification of *Contingency* relations as *Temporal*, often in cases such as (1)[2] where the facticity of the sentence cannot be judged without context:

(1) But **when** market interest rates move up rapidly, increases in bank CD yields sometimes lag.

Among the second-level relations, performance on most relations is generally good, with most frequent relations having an F-measure of more than 0.9, but several relations are frequently misidentified: The distinction between *Concession* and *Contrast* – obviously a relatively central one, which however depends on the semantic content of the connective arguments – cannot always be made reliably, and *Concession* as the less frequent relation shows low precision and recall. Within the *Expansion* relations, the lower-frequency relations (of which only *List* is shown) are never predicted because the features used are not strong enough to overcome the preference for the more frequent relations. Within the *Temporal* relations, we see that the effect of misclassifications such as in example (1) is more predominant on the *Temporal.Synchrony* relations.

Pitler and Nenkova's features use the Penn Treebank in its original form, including, on one hand, traces, and, on the other hand, function labels which indicate temporal (−TMP), purpose (−PRP) or other adverbial modification (−ADV).[3] Given an automatic parse, this information would have to be reconstructed, since parsing models are always trained on a version of the treebank that has traces and such semantic function labels removed.[4] Furthermore, the reconstruction of traces and function labels is somewhat error-prone (Gabbard et al., 2006 report an F-measure of about 85% for semantic function tags, and 75% for traces) which means that using this information in sense prediction is prone to overestimating the actual performance in a complete system.

To quantify the influence of this additional gold-standard information, we compute a variant of the syntax features where the trace feature is not used and function labels are stripped from the nodes. As can be seen in Table 5, this version of the syntactic features gives results that are very close to the results that one gets with only the string of the connective.

While the inclusion of tense information cannot improve over the information contained in the semantic function tags (see the *conn+syntax*^{A} and *conn+syntax*^{A}*+tense* rows in Table 5), the incorporation of tense/mood information on the heuristically determined ARG1 (if present in the same sentence) yields useful results by itself.

In contrast, including a single feature that summarizes the syntactic environment (subordinating coordinator, clausal PP, sentence coordination, etc.) and tense features *for the modifiee (Arg1) only* yields results that are close to those with semantic function tags.

[2]The example is annotated as *Contingency.Condition.Hypothetical*, but predicted as *Temporal.Synchrony*

[3]It is not clear from Pitler and Nenkova's paper whether they used a version with function labels or without, since they do not mention it; as they use traces, the most plausible interpretation is that they used a version where function labels are intact.

[4]The −TMP function label on noun phrases is usually kept, since it reflects a syntactic distinction – adverbial versus argument role of the NP – rather than a semantic one and is useful for the parser itself.

	d=1	d=2	d=3
hierarchical			
connective only	0.946	0.839	0.790
conn+syntax^A	0.954	0.847	0.796
conn+syntax^B	0.945	0.840	0.788
w/traces	0.948	0.843	0.792
w/function tags	0.954	0.847	0.796
conn+verb(arg1)	0.952	0.845	0.798
conn+syn^B+pos(arg1)	0.949	0.843	0.794
conn+pos(both)	0.949	0.843	0.794
conn+syn^B+pos(both)	0.947	0.839	0.788
greedy			
connective only	0.946	0.840	0.792
conn+syntax^A	0.955	0.847	0.798
conn+verb(arg1)	0.953	0.845	0.800

syntax^A: with traces and function tags syntax^B: without traces or function tags

Table 5: Different versions of syntactic and tense/mood features

Both the results for syntax including semantic function tags and those for the inclusion of Arg1-related verb features yield improvements over the connective-only version that are statistically significant according to a paired t-test. (All are significant at the $p < 0.05$ level; the improvements on the first level yield p-values around 10^{-5}).

Tables 6 and 7 summarize the behavior of relation prediction over several connectives. Besides the fact that the more difficult task of distinguishing relations according to the larger set also leads to more connectives showing ambiguities, we see that, firstly, the distinction between topicalized and non-topicalized adjunct clauses (summarized as a feature indicating whether the connective is at the start of a sentence, in the column named *conn+first*) has relatively limited benefits. Secondly, the actual syntactic features (*conn+syn^B*, without semantic function labels) and the tense/mood-based features (*conn+mood*) are useful in the case of different connectives – "*since*", for example, does not benefit much from syntactic features but shows a strong improvement when tense and mood information is added.

4 Conclusion

In this paper, we presented first results on the classification of discourse relations using a novel approach that makes use of the hierarchical structure of the label set of the Penn Discourse treebank, and provided an error analysis that extends to the lower levels of

connective	frequency	conn	conn+first	conn+syn^B	conn+syn^A	conn+verb(arg1)
since	154	0.571	0.571	0.675	**0.935**	0.909
finally	30	0.633	0.933	0.867	0.867	**0.933**
in turn	27	0.704	0.704	0.704	0.704	0.704
even as	11	**0.727**	0.636	0.364	0.455	0.636
while	652	0.729	0.727	0.729	**0.839**	0.805
as	588	0.786	0.786	0.781	**0.810**	0.781
as long as	20	**0.800**	0.786	0.750	0.700	0.750

Connectives that occur at least 10 times and have at most 80% accuracy

Table 6: Ambiguous connectives at the coarsest level

connective	frequency	conn	c+first	c+syn^B	c+syn^A	c+arg1	c+syn^B+arg1
rather	14	0.286	**0.643**	0.429	0.357	**0.643**	0.500
as soon as	17	0.294	0.294	0.294	0.176	**0.412**	0.176
nevertheless	30	0.300	**0.533**	0.300	0.333	0.300	0.333
in fact	70	0.300	0.386	0.286	0.300	0.343	**0.429**
finally	30	0.367	**0.667**	0.633	0.533	**0.667**	**0.667**
although	277	0.498	0.588	0.520	0.549	0.592	**0.606**
still	156	0.500	0.429	0.462	**0.506**	0.417	0.449
since	154	0.571	0.571	0.669	**0.929**	0.903	0.896
though	187	0.588	**0.652**	0.540	0.551	**0.652**	**0.652**
while	652	0.598	0.598	0.604	**0.718**	0.667	0.672
indeed	86	**0.605**	0.593	0.593	0.593	0.570	0.558
when	837	**0.611**	0.608	0.609	0.609	0.596	0.588
in particular	13	**0.615**	0.538	0.538	0.538	0.538	0.538
yet	88	**0.648**	**0.648**	0.523	0.432	0.523	0.545
overall	10	**0.700**	0.600	0.400	0.300	0.600	0.400
in turn	27	0.704	0.704	0.704	0.704	0.704	0.704
even as	11	**0.727**	0.636	0.182	0.273	0.636	0.273
as	588	0.745	0.745	0.736	**0.767**	0.743	0.745
in the meantime	12	0.750	0.833	**1.000**	**1.000**	0.833	**1.000**
but	2767	**0.790**	**0.790**	0.789	0.788	0.789	0.785
nor	24	**0.792**	**0.792**	0.667	0.667	0.750	0.667
meanwhile	160	**0.800**	**0.800**	**0.800**	0.775	0.794	0.794
as long as	20	**0.800**	0.750	0.750	0.700	0.750	0.750
ultimately	17	**0.882**	0.765	0.706	0.647	0.765	0.706
now that	20	0.900	0.900	0.550	**0.901**	0.900	0.750

Connectives that occur at least 10 times and have at most 80% accuracy

Table 7: Ambiguous connectives at the medium level

the label hierarchy.

Considering the purpose of applying discourse tagging to raw text, it would be desirable to achieve the tagging of connectives at the granularity of second-level, rather than top-level categories in the Penn Discourse Treebank's inventory, since many important distinctions (*Contrast* versus *Concession*, or *Cause* versus *Condition*) are only made at the second level of the taxonomy. For many of these finer distinctions, neither Pitler and Nenkova's syntactic features nor the tense/mood based that we presented here are sufficient to reach high (>90%) accuracies, despite Pitler and Nenkova's encouraging results on the coarser top-level relation categories.

One plausible reason for this is that the shallow information used by current approaches is not sufficient to reproduce the more semantic distinctions on the finer levels of the taxonomy. Another plausible reason, which has also been pointed out by Pitler and Nenkova concerning the coarser-level distinctions and which we cannot exclude at this point, would be that system accuracy is bounded by annotator agreement: Some distinctions among those in the Penn Discourse Treebank are hard to make reliably even for humans, and similarly our results come close to the levels of annotator agreement reported by Prasad et al. (2008) for the PDTB – 84.5% for the second level, against 84% agreement, and 79.5% for the third level, compared to an agreement figure of 80%.

Our evaluation on the finer levels of the relation taxonomy, however, is slightly more lenient than the annotation in the Penn Discourse Treebank: as we allow any subtype for the underspecified relations where annotators disagreed on the finer relations, these disagreement cases are mostly counted as correct, whereas counting a more specific label as wrong (which would mean that the majority of such disagreement cases would be counted as a disagreement between system and gold annotation, since the system only very rarely assigns an underspecified label) would yield markedly lower results of about 68% for the third-level relations, which would allow for hope of further improvement through more semantic features.

References

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Massimiliano Ciaramita, Thomas Hofmann, and Mark Johnson. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, 2003.

Robert Elwell and Jason Baldridge. Discourse connective argument identification with connective specific rankers. In *Proceedings of ICSC-2008*, 2008.

Ryan Gabbard, Mitchell Marcus, and Seth Kulick. Fully parsing the Penn Treebank. In *HLT/NAACL 2006*, 2006.

Barry Haddow. Acquiring a disambiguation model for discourse connectives. Master's thesis, School of Informatics, University of Edinburgh, 2005.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *EMNLP 2009*, 2009.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In *TLT 2005*, 2005.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. A sequencing model for situation entity classification. In *ACL 2007*, 2007.

Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *ACL 2009 short papers*, 2009.

Emily Pitler, Annie Lous, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *ACL-IJCNLP 2009*, 2009.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-2003)*, 2003.

Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416, 2008.

Building a Discourse-Annotated Dutch Text Corpus

Nynke van der Vliet[*], *Ildikó Berzlánovich*[*], *Gosse Bouma*[*], *Markus Egg*[†] *and Gisela Redeker*[*]
* University of Groningen, †Humboldt University, Berlin.

Abstract
We are compiling a corpus of Dutch texts annotated with discourse structure and lexical cohesion, containing initially 80 texts from expository and persuasive genres. We are using this resource for corpus-based studies of discourse relations, discourse markers, cohesion, and genre differences. We are also exploring the possibilities of automatic text segmentation and semi-automatic discourse annotation. This paper discusses our design choices in text selection and segmentation and in the annotation of discourse structure and lexical cohesion.

1 Introduction

Discourse researchers from descriptive, cognitive, formal, and computational backgrounds unanimously subscribe to the view that texts are structured entities that exhibit coherence and cohesion (for a recent overview see Taboada and Mann (2006b)). Coherence refers to the way sentences or utterances combine to convey the informational and intentional (e.g., expressive or persuasive) meanings of the text. Cohesion refers to elements (conjunctions and other so-called "cue phrases") that signal how utterances or larger text parts are related to each other, and to the way lexical elements like pronouns and definite noun phrases refer back to other items in the discourse (Halliday and Hasan, 1976). The main goal of our corpus-building effort is to provide the basis for investigating discourse structure, relational and lexical cohesion, and their interactions with genre, i.e., to support the modeling of textual organization.

Much of the theoretical and empirical research on relational coherence has focused on local coherence relations and their linguistic signaling (e.g., Sanders et al. (1992, 1993); Knott and Sanders (1998); Webber et al. (2003), Prasad et al. (2008)). Configurational issues concerning the hierarchical composition of larger stretches of text that arise from recursive application of coherence relations, have received some attention in computational linguistics, but lack a substantial empirical foundation. Various structures have been proposed, in particular, binary trees (e.g., Carlson et al. (2002); Stede (2004)), *n*-ary trees (e.g., Mann and Thompson (1988); Webber (2004), Polanyi et al. (2004); Thione et al. (2004)), and less constrained graph structures (Danlos (2004); Wolf and Gibson (2005)).

The interplay of relational discourse structure with referential and lexical cohesion has been investigated with a focus on the use and interpretation of anaphoric expressions (Fox (1987); Grosz et al. (1995); Kehler (2002); Poesio et al. (2004)); much less attention has been devoted to the role of lexical cohesion in co-determining the overall textual organization (but see Hasan (1984) and Hoey (1991)).

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena", pages 157-171, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

Textual organization cannot be studied without consideration of the variability between text genres (see, e.g., Eggins and Martin (1997), Webber (2009)). In particular, some texts are organized around a central purpose, e.g. a claim that is argued for or a request or proposal the text is intended to support, while descriptive or expository texts are usually organized around a central theme, moving through sub-themes or aspects. This difference is relevant for both, the relational structure and the role of lexical cohesion. The corpus therefore covers a range of genres.

By annotating relational and lexical organization in a variety of text types, this project will create a Dutch language resource for corpus-based discourse research, computational modeling, and applications like question answering and summarization.

2 Corpus design

Our aim is to provide a reliably annotated "gold standard" resource covering a range of genres. The emphasis on quality and richness of the manual annotation limits the size of the corpus, as careful annotation work is extremely time consuming.

2.1 Text selection

The corpus covers a range of text genres, including, in particular, expository texts, whose main purpose is to present information to the reader, and persuasive texts that aim to affect the readers intentions or actions. The texts vary in length between a minimum of approximately 190 words and a maximum of approximately 400 words. Longer texts become unwieldy for relational analysis, and top-level relations tend to be rather uninformative juxtapositions (Taboada and Mann, 2006b).

The corpus consists of 40 expository texts and 40 persuasive texts. For the expository subcorpus, 20 texts have been selected from online encyclopedias on astronomy[1] and 20 from a popular scientific news website[2]. The persuasive texts are 20 fundraising letters from humanitarian organizations and 20 commercial advertisements from lifestyle and news magazines.

Encyclopedia entries as well as popular scientific news are learned exposition, i.e., texts that are strictly informational in purpose, but moderately technical in content and style, and that take the general public as their audience. In this way, we excluded scientific exposition, which is more abstract and technical in style and targets professional scientific audience (e.g., academic prose). Fundraising letters and advertisements are prototypical persuasive genres that have received much attention in the literature (e.g., Bhatia (1998), Kamalski (2007)). They have a clear and focused purpose and are directed at a general audience.

[1] http://www.astronomie.nl; http://www.sterrenwacht-mercurius.nl/encyclopedie.php5
[2] http://www.scientias.nl/category/astronomie

2.2 Annotation

The starting point of our annotation work is a syntax-based segmentation of the texts into clausal atomic units, which has been developed in an extended training phase involving consistency checking aided by a collection of examples (see section 3 below). We then add annotations for discourse structure, relational cohesion, and lexical cohesion, which we are briefly introducing here (for details see sections 4 and 5).

For the analysis of relational discourse structures, we chose the widely used Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Taboada and Mann, 2006a) in its "extended classic" variant. The XML annotation is created using O'Donnell's RSTTool[3] (O'Donnell, 1997). The definitions of the RST relations are available from the RST website[4].

Previous research has shown how combining genre analysis and RST analysis enriches our understanding of discourse structure (e.g., Taboada and Lavid (2003), Gruber and Muntigl (2005)). We are therefore overlaying the RST-trees with a segmentation of the global text units according to the genre-specific *moves* they realize (Upton and Cohen, 2009). The mapping of the sequence of moves onto the RST-trees adds relational and hierarchical information.

Three subsystems of cohesion contribute to the organization of a text: relational cohesion (lexical or phrasal elements that signal coherence relations), referential cohesion (anaphoric chains, spatial/temporal chaining and ellipsis), and lexical cohesion arising from the semantic network of the lexical items in the text (Halliday and Hasan (1976); Halliday and Matthiessen (2004)). In this project we focus on relational cohesion and lexical cohesion.

The analysis of relational cohesion will include all lexical or phrasal elements (discourse markers) in the text that signal coherence relations at local and global levels of discourse. We are currently developing our methodology for this analysis.

The analysis of lexical cohesion starts by identifying all content words (nouns, verbs, adjectives, adverbs) and then locating their neighboring lexical associates in other discourse units. The XML annotation is created with an MMAX-based tool (Müller and Strube, 2001).

All annotations are done separately by at least two annotators and then discussed. Inter-annotator agreement using Kappa shows a high level of agreement on the segmentation: .97 for the encyclopedia texts and .99 for the fundraising letters. We computed inter-annotator agreement for the RST analysis for two fundraising letters and two encyclopedia texts, using the methods proposed in Marcu et al. (1999). On average, the agreement was .88 on the spans and .82 on the nuclearity. The agreement on the RST relation labels was only .57. We suspect (and hope to confirm with the complete data set) that this is not a general deficiency of our annotation but a problem that can mainly

[3] available from `http://www.wagsoft.com/RSTTool/`.

[4] `http://www.sfu.ca/rst/`

be attributed to a few rather confusable relations such as *Joint* versus *Conjunction*. As Marcu et al. (1999) point out, these Kappa values are comparable with the agreement in other corpora.[5]

The annotation of all 80 texts in the core corpus will be complete by March 2011. Manuals detailing the segmentation and annotation rules will be made available along with the corpus.

3 Segmentation

An essential step in discourse analysis is the identification of suitable Elementary Discourse Units (EDUs). Various definitions of EDUs exist, ranging from a fine-grained segmentation to segmentation at sentence level. In classic Rhetorical Structure Theory (RST), clauses are considered to be EDUs, except for subject and object clauses, complement clauses, and restrictive relative clauses (Mann and Thompson, 1988).

For the annotation of the RST Discourse Tree Bank, Carlson and Marcu (2001) use a fine-grained segmentation in which they also treat complements of attribution verbs and phrases that begin with a strong discourse marker (e.g. *because of, in spite of, according to*) as separate EDUs. Relative clauses, nominal postmodifiers, or clauses that break up other legitimate EDUs are treated as embedded discourse units. Based on this, Lüngen et al. (2006) developed segmentation guidelines for German text, but in contrast to Carlson and Marcu (2001) they exclude restrictive relative clauses, conditional clauses, and proportional clauses (clauses combined by comparative connectives). Grabski and Stede (2006) suggest to also include prepositional phrases as EDUs. Tofiloski et al. (2009) adhere more closely to the original RST proposals (Mann and Thompson, 1988) and segment coordinated clauses, adjunct clauses and non-restrictive relative clauses. To our mind, these differences follow from attempts to include semantic considerations in the definition of EDUs (i.e., including at least some proposition-denoting yet non-clausal segments among the EDUs).

For Dutch, as far as we know, such an elaborate investigation of what should count as an EDU has not yet been done. RST annotations of Dutch text have used the segmentation of the original RST proposals (Abelen et al., 1993) or taken clauses containing a finite verb (den Ouden et al., 1998) or whole sentences (Timmerman, 2007) as EDUs.

3.1 Segmentation principles

The segmentation we use for the Dutch corpus is fairly coarse. The EDUs are independent or subordinate clauses or other complete utterances (independent fragments). The definition of an elementary discourse unit is guided by the question of whether a discourse relation could hold between the unit and another segment. EDUs are typically

[5] Brown corpus (Francis and Kucera, 1979), MUC corpus (Chinchor, 2001), WSJ corpus (Carlson et al., 2002)

propositions or segments that constitute speech acts of their own. The segmentation principles are based on syntax and punctuation rather than semantic criteria.

Like Tofiloski et al. (2009), we treat simple sentences (1), coordinated clauses (2), subordinate clauses (3) and non-restrictive relative clauses (4) as EDUs.

(1) [Elke donatie is waardevol!]
[Each donation is valuable!]

(2) [Cavine kreeg aidsremmers][en dat maakte een levensgroot verschil.]
[Cavine got aids medication][and that made a huge difference.]

(3) [Omdat de EU binnenkort beslist over nieuwe regels,][voeren we de druk op de politiek nu hoog op]
[Because the EU will decide on new regulations soon][we are now strongly increasing our pressure on politics.]

(4) [Dit gat wordt veroorzaakt door een van de maantjes van Saturnus, Mimas,][die de ringen verstoort.]
[This gap is caused by one of the moons of Saturn, Mimas,][which disturbs the rings.]

In contrast to Tofiloski et al. (2009), we consider coordinated elliptical clauses (i.e. clauses that share a verb that is elided in one of the clauses, as in (5)) as separate EDUs, because the two clauses that share a verb can be seen as two separate predicates. This also applies to clauses that share a noun phrase as subject, as in (6). In Carlson and Marcu (2001), clauses with an ellipted subject are segmented as EDUs as well, whereas clauses with an ellipted verb are only treated as EDUs when there are strong rhetorical cues marking the discourse structure as in (7)[6].

(5) [De planeet draait in 58.6 dagen om haar as] [en in 88.0 dagen om de zon.]
[The planet turns around its axis in 58.6 days][and around the sun in 88.0 days.]

(6) [De operatie duurde 15 minuten][en kostte 35 euro.]
[The surgery took 15 minutes][and cost 35 euros.]

(7) [Back then, Mr. Pinter was *not only* the angry young playwright,] [*but also* the first] [to use silence and sentence fragments and menacing stares, almost to the exclusion] [of what we preciously understood to be theatrical dialog.] (wsj_1936)

Non-restrictive relative clauses as in (8) and embedded clauses between parentheses as in (9) are considered to be embedded discourse units. Restrictive relative clauses, subject and object clauses, and complement clauses are not treated as separate EDUs (following classic RST). Contrary to Carlson and Marcu (2001), Lüngen et al. (2006),

[6]Example from Carlson and Marcu (2001)

and Jasinskaja et al. (2007), we do not recognize non-clausal appositives as in (10) as separate EDUs.

(8) [Echter gedurende de nacht, [die op Mercurius maanden lang kan duren,] daalt de temperatuur tot zo'n -185 graden Celsius.]
[However during the night, [which can last for months on Mercury,] the temperature decreases to about -185 degrees Celsius.]

(9) [De binnenste maan [(van 2002 tot 2005 is dat Epithemeus)] beweegt iets sneller dan de buitenste] [en haalt die ander langzaam (met 450 meter per minuut) in.]
The innermost moon [(from 2002 to 2005 this is Epithemeus)] moves a bit faster than the outermost] [and slowly (with 450 meters per minute) catches up with the other.]

(10) [Het tweede type terrein, het laagland, telt relatief nog minder kraters dan het hoogland.]
[The second terrain type, the lowland, contains even fewer craters than the highland.]

Our segmentation uses punctuation in connection with syntax. Periods, exclamation marks and question marks are EDU boundaries, except for periods that are used in abbreviations, acronyms, dates and so forth. Independent fragments (subclausal expressions ending with a period) as in (11) are considered to be EDUs.

(11) [Leuke hebbedingetjes.]
[Nice gadgets.]

Colon or semicolon are only treated as separation markers when the subsequent material is a clause as in (12). If it is a non-clausal expression, as in (13), it is not segmented. The same rule applies for text structures between hyphens or parentheses: clauses as in (9) or participle structures as in (14) are segmented as EDUs, but non-clausal material as in (15) is not segmented.

(12) [Daar knapt ze zichtbaar van op;][ze begint ook weer te praten!]
[From that, she recuperates visibly;][she even starts to talk again!]

(13) [In 2005 zijn nog twee maantjes van Pluto ontdekt: Nix en Hydra.]
[In 2005, two more small moons of Pluto were discovered: Nix and Hydra.]

(14) [Wat er binnen deze bol [(horizon genoemd)] gebeurt weten we niet.]
[What happens inside this globe [(called horizon)] we don't know.]

(15) [De krater Pan (inslagkrater), de grootste krater, is 100 kilometer in doorsnede] [en minstens 8 kilometer diep.]
[The crater Pan (impact crater), the biggest crater, is 100 kilometers in diameter][and at least 8 kilometers deep.]

4 Discourse structure

The annotation of discourse structure is intended to capture the hierarchical structures arising from coherence relations between discourse units, but also the genre-specific structures that can help in understanding genre differences in discourse structure.

4.1 Rhetorical Structure Theory

There is wide agreement that discourse is hierarchically structured, and many current theories assume that this structure arises from the recursive application of coherence relations. Discourse-annotated corpora are particularly useful for investigating the realizations, linguistic marking, and genre-specific uses of coherence relations (e.g., Webber (2009); Taboada et al. (2009); see also the discussion in Taboada and Mann (2006a,b)) and we are researching such questions with our corpus. In addition, however, we are also interested in the configurational characteristics of discourse structure. We thus differ from annotation efforts like the Penn Discourse TreeBank (Prasad et al., 2008) that focus mainly on coherence relations and on implicit and explicit connectives. For us, it is essential to represent the full hierarchical structure of our texts.

Rhetorical Structure Theory (RST; Mann and Thompson (1988)) has proven successful for the analysis of whole texts and has been widely applied (for an overview see Taboada and Mann (2006a,b)) to texts of various languages and used for the annotation of large text corpora (Carlson et al. (2002), Stede (2004)).

We base our analyses on the set of 30 relations as defined in "extended classic" RST. We do not follow Carlson and Marcu (2001), who use a much larger set of relation labels (mostly necessitated by their more fine-grained segmentation) (for a critical discussion of both variants of RST, see Stede (2008)).

In particular, we do not use Carlson and Marcu's (2001) *Attribution* and *Same* relations, which we consider problematic. *Attribution* is defined in Carlson and Marcu (2001) as the relation between a direct or indirect quotation and its attributing phrase or clause. This relation is arguably of a categorically different kind than coherence relations (Tofiloski et al. (2009), Skadhauge and Hardt (2005)). In classic RST, complement clauses and speech parentheticals are not considered as separate EDUs. This means that speech-reporting EDUs can enter coherence relations as speech events or by virtue of the speech that is reported (in particular when the quotation is continued in subsequent EDUs). This flexibility fits in well with the idea that semantic relations in discourse are often underspecified (Egg and Redeker, 2008).

The pseudo-relation *Same* is introduced by Carlson and Marcu (2001) to link two discontinuous parts of an EDU that is interrupted by another, parenthetically embedded, EDU. In classic RST, parenthetical EDUs are extracted and placed after their host EDU, thus obviating the need for a pseudo-relation (see, e.g., Redeker and Egg (2006)).[7]

[7] Borisova and Redeker (2010) point out problems involving the *Same* relation in the Discourse GraphBank (Wolf et al. (2003)).

4.1.1 Discourse trees or graphs?

Rhetorical Structure Theory assumes that the discourse structure of a text can be represented as an ordered tree. In this tree all text parts are in some way connected to the root of the tree, the most central text part. However, it has been claimed that tree structures are not sufficient to represent discourse structure (Asher (2008); Lee et al. (2008); Wolf and Gibson (2005)). Wolf and Gibson (2005) show that crossed dependencies (i.e. structures in which discourse units ABCD (not necessarily adjacent) have relations AC and BD) and multiple-parent structures (where a unit enters more than one coherence relation and is thus dominated by more than one node) occur abundantly in their Discourse GraphBank (Wolf et al. (2003)). They argue that these constellations, which violate the tree-structure constraints, are necessary to describe the text structures in their corpus, and that a more complex graph structure is thus required to represent the discourse structure of a text.

Webber (2006) and Egg and Redeker (2008, 2010), however, argue that the chain graphs in the Discourse GraphBank conflate discourse constituency and anaphoric dependency. Egg and Redeker (2008) point out that the analyses discussed in Wolf and Gibson (2005) have plausible tree-based alternatives and Egg and Redeker (2010) further support this argument with data from the Discourse GraphBank. While this question is not yet settled, we do find that trees are adequate data structures to represent the constituent structure of discourse for the texts in our corpus and thus use RST-trees to annotate discourse structures.

4.1.2 Non-binary trees

Given the assumption that discourse structure can be adequately represented by trees, it is tempting to consider the still stronger assumption that would only allow binary trees, which are much simpler and computationally more tractable. This restriction is indeed often implemented in discourse parsers (e.g. Marcu (2000); Soricut and Marcu (2003); Reitter (2003)). In our project, we choose plausibility and validity of our analyses over computational tractability and allow non-binary structures in our RST trees.

RST-trees do contain mostly binary relations (in particular the asymmetric *nucleus-satellite* relations),[8] but they also admit non-binary structures with multiple nuclei or multiple satellites relating to one nucleus. In the first case, several nuclei are involved in one *multinuclear* relation, e.g., *List*, *Sequence* or *Joint*. Binary representations of such structures (proposed, e.g., by Egg and Redeker (2008)) involve a stacking of binary relations, implying a hierarchical ordering (left- or right-branching or pairwise clustering) among the list constituents. These binary representations do not reflect the

[8] RST distinguishes two kinds of relations: The asymmetric *mononuclear* relations like *Elaboration* or *Justify* relate a *nucleus* (centrally important) and a *satellite* (additional information, which could in many cases be left out without rendering the text incoherent). The symmetric *multinuclear* relations like *List* or *Joint* relate discourse entitites of equal status.

equal importance of the items in the multinuclear relation.

In the second kind of non-binary structures, several *nucleus-satellite* relations share the same nucleus, e.g., when the central request of a fundraising letter is supported by various preceding or succeeding *Motivation* and/or *Justify* satellites, as described in Abelen et al. (1993), or when several separate *Elaborations* provide details about the contents of one nucleus. A binary representation of these structures requires that one of the satellites of the shared nucleus is included in the nucleus of another satellite, which is in many cases not plausible.[9]

We consider the regular occurrence of non-binary structures sufficient reason to assume that discourse structure representations require non-binary trees.

4.2 Moves

For comparisons of the global text structure across genres, we identify the genre-specific major building blocks of the texts using *move analysis* (Upton and Cohen, 2009). We identify the functional components, so-called moves, in the text. A move is realized by at least one EDU. Contrary to, e.g. Biber et al. (2007), we do not recognize moves below EDU level and do not allow embedding of moves. The moves in our analysis create a linear, non-hierarchical partition of the EDUs in the text. Each genre has a particular set of move types that occur regularly in texts of that genre. Some move types are obligatory. Any move type may be realized more than once in a particular text. In the encyclopedia entries, we identify the move types *name, define* and *describe*. For the fundraising letters, we follow Upton (2002), who identified seven move types labeled *get attention, introduce the cause and/or establish credentials of organization, solicit response, offer incentive, reference insert, express gratitude, conclude with pleasantries*. The move structure of advertisements is based on Bhatia (2005) and contains the following move types: *get attention, justify the product or service by establishing a niche, detail the product or service, establish credentials, endorsement/testimonial, offer incentive, use pressure tactics, solicit response, and reference to external material*. Finally, the starting point for determining the move structure of the popular scientific news will be van Dijk's superstructure of news (van Dijk, 1988), which is a hierarchical structure containing the main genre elements of news in general.

5 Cohesion

Parallel to the discourse structure annotation, we are annotating the corpus for relational cohesion and lexical cohesion.

[9] An alternative explanation that first collects all satellites in a *List* or *Joint* segment, which then as a whole functions as the sole satellite of the respective nucleus is only feasible in a subgroup of these cases, in which all satellites occur on the same side of the nucleus (before or after it) and are related to the nucleus in terms of the same relation.

5.1 Relational cohesion

Relational cohesion concerns the lexical or phrasal elements (*discourse markers*) in a text that signal coherence relations, both at the local and global levels of discourse. Some relations are often signaled by discourse markers, e.g. the conjunction relation (*and, also*), but others are implicit and do not contain clear cues (Taboada, 2006).

In a pilot study we have analyzed the distribution and explicit signaling of coherence relations in 20 encyclopedia entries and 20 fundraising letters. Intra-sentential relations are much more often signaled than inter-sentential relations (69% vs 16%), presumably reflecting the fact that intra-sentential clause combining usually involves an obligatory conjunction or adverb, while there is no such syntactic requirement for marking inter-sentential relations.

Future work will include the annotation of discourse markers (conjunctions and conjunctive adverbs) and their scopes, comparable to the annotations in the Penn Discourse Treebank (Prasad et al., 2008), with the dual aim of theoretical investigations and the development of a semi-automatic parsing tool for coherence relations.

5.2 Lexical cohesion

In our analysis of lexical cohesion, we aim to cover all types of semantic relations among lexical items in the text (see section 5.2.2 below; for recent work on an overview of approaches to lexical cohesion, see Tanskanen (2006)). We include only relations across elementary discourse units (EDUs), not within EDUs. This allows us to investigate the alignment between discourse structure and lexical cohesion, as both structures are based on the same units. At a finer level, we also study the co-occurrence of lexical cohesion types with coherence relations.

5.2.1 Selection of lexical items

As we are interested in the contribution of *lexical* cohesive relations, we exclude pronouns and do not follow referential chains through the text. The class of items for participating in lexical cohesion includes content words (nouns, verbs, adjectives, and adverbs of place, time, and frequency) and proper names. Proper names are treated as one unit. The elements of multi-word units (except for proper names) are treated as separate lexical items, while compounds are taken as indecomposable single units.

5.2.2 Categories of lexical cohesive relations

The categories we distinguish for lexical cohesive relations are listed in Table 1. By *repetition* we mean word repetition. The lexical items in full repetition have fully identical word form or they differ only in their inflectional suffix, whereas lexical items in partial repetition have different derivational suffixes in their word form. Under the heading

Category		Example
Repetition	Full repetition	*planet - planet*
	Partial repetition	*planet - planetary*
Systematic semantic relations	Hyponymy	*sun - star*
	Hyperonymy	*gas - hydrogen*
	Co-hyponymy	*Venus - Mercury*
	Meronymy	*planet - solar system*
	Holonymy	*solar system - sun*
	Co-meronymy	*Earth - sun*
	Synonymy	*life - existence*
	Antonymy	*light - heavy*
Collocation		*light - star*

Table 1: Categories of lexical cohesion

systematic semantic relations we include the traditional lexical semantic relations. The lexical cohesive relation *collocation* is formed between two lexical items which tend to occur in similar lexical environments because they describe things that tend to occur in similar situations or contexts in the world (Morris and Hirst, 1991). Note that this use of the term implies a meaning relation between the lexical items in contrast to its use in corpus linguistics, where collocation refers to the mere co-occurrence of words (Stubbs, 2001), which is not a sufficient criterion for lexical cohesion.

We identify relations arising from lexical meaning (e.g., *planet - Earth*) and ignore accidental meaning relations that arise from context. In addition, we identify relations that are easy for the reader to identify with general background knowledge and for which no further knowledge or textual context is necessary for their identification (e.g., we identify the relation of *astronomer* with *Kepler*, but not with *Richard Walker*, although the textual context helps us understand that Richard Walker is also an astronomer). This question is strongly related to the issue of register-sensitive and domain-sensitive relations. Although we aim to identify general relations, i.e., relations which are not specific of a certain register or domain, the annotators have to face the difficulties of drawing the line between general and context-dependent.

5.2.3 Lexical cohesion links as a graph structure

Lexical cohesive links build up graph structures in the text. In our analysis any candidate item can enter into a lexical cohesive relation with any other candidate items as long as there is a meaning relation between them. For each lexical item in a text, we identify its lexical links—if any—to preceding lexical items (lemmas), ignoring any links among the words inside an EDU. If a lexical item is linked to more than one preceding item, all of those relations are registered as cohesive links. Similarly, if a lexical item enters into cohesive relations with more than one item occurring in succeeding EDUs, all those links are counted.

In this way, we build up networks that represent the lexical cohesive structure of a text. By assigning graph structures to lexical cohesion, we differ from previous studies that identified lexical cohesive chains in text (e.g., Hasan (1984), Morris and Hirst (1991)) and follow those that identify networks (Hoey, 1991). Modeling lexical cohesion with graph structures provides a much richer representation than the lexical cohesive chains model. It also allows us to measure the centrality of a lexical item by its centrality in the network.

6 Conclusion

The resource we are building aims at a high standard of empirical validity (very careful annotation based on detailed, explicit rules) and coverage across a theoretically motivated selection of text genres. With a core of 80 texts, the corpus is rather small for computational applications, but still large enough for distributional analyses and structural comparisons.

We have been using the initially completed parts of this corpus to investigate genre differences in the use of discourse relations and in the occurrence of lexical cohesion relations and the interaction of these two aspects of textual organization (Berzlánovich and Redeker, 2011). As our discourse structure annotation follows the widely used "classic" RST, we expect our corpus to support cross-linguistic research through its comparability with RST-based corpora in other languages.

Our segmentation rules are surface oriented (based on syntax and punctuation) and have been implemented in an automatic segmenter (van der Vliet, 2010). Future work will include the annotation of discourse markers with the dual aim of theoretical investigations and the development of a semi-automatic parsing tool for coherence relations. With an eye on crosslinguistic research on discourse and discourse markers in the spirit of Knott and Sanders (1998), we will strive for compatibility with the annotation in the Penn Discourse TreeBank (Prasad et al., 2008), but will more freely allow markers to signal global coherence relations among larger text spans (which is discouraged by PDTB's *Minimality Principle* (Prasad et al. (2007): 19), according to which annotators have to select the minimally necessary segments).

Finally, we also envisage combining our lexical cohesion analysis with computational coreference resolution (Hendrickx et al., 2008) and testing our network model of lexical cohesion against approaches based on lexical chaining (see, e.g., Barzilay and Elhadad (1997)).

Acknowledgments

The work reported here is supported by grant 360-70-280 of the Netherlands Organization for Scientific Research (NWO). For online documentation of the program *Modeling discourse organization* see `www.let.rug.nl/mto`. We are grateful to three anonymous reviewers for their valuable comments on an earlier version of this paper.

References

Eric Abelen, Gisela Redeker, and Sandra A. Thompson. The rhetorical structure of US-American and Dutch fund-raising letters. *Text*, 13(3):323–350, 1993.

Nicholas Asher. Troubles on the right frontier. In Peter Kühnlein and Anton Benz, editors, *Constraints in Discourse*. Benjamins, Amsterdam, 2008.

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 17. Madrid, Spain, 1997.

Ildikó Berzlánovich and Gisela Redeker. Genre-dependent interaction of coherence and lexical cohesion in written discourse. In *Corpus Linguistics and Linguistic Theory*, 2011. To appear.

Vijay K. Bhatia. Generic patterns in fundraising discourse. *New directions for philanthropic fundraising*, (22):95–110, 1998.

Vijay K. Bhatia. Generic patterns in promotional discourse. In Helana Halmari and Tuija Virtanen, editors, *Persuasion across genres: A linguistic approach*, pages 213–228. Benjamins, Amsterdam, 2005.

Douglas Biber, Ulla Connor, and Thomas A. Upton. *Discourse on the move: Using corpus analysis to describe discourse structure*. Benjamins, Amsterdam, 2007.

Irina Borisova and Gisela Redeker. Same and Elaboration relations in the Discourse Graphbank. In *Proceedings of the 11th annual SIGdial Meeting on Discourse and Dialogue, Tokyo*, 2010.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 2001.

Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. RST Discourse Treebank. *Linguistic Data Consortium*, 2002.

Nancy Chinchor. *Message Understanding Conference (MUC) 7*. Linguistic Data Consortium, Philadelphia, 2001.

Laurence Danlos. Discourse dependency structures as constrained DAGs. In M. Strube and C. Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, Massachusetts, USA*, pages 127–135, 2004.

Hanny J .N. den Ouden, Carel H. van Wijk, Jacques M.B. Terken, and Leo .G.M. Noordman. Reliability of discourse structure annotation. *IPO Annual Progress Report*, 33:129–138, 1998.

Markus Egg and Gisela Redeker. Underspecified discourse representation. In P. Kühnlein and A. Benz, editors, *Constraints in Discourse (CID), Dortmund, June 3-5, 2005*, pages 117–138. Benjamins, Amsterdam, 2008.

Markus Egg and Gisela Redeker. How complex is discourse structure? In *Proceedings of LREC'10, Malta, 17-23 May 2010*, pages 1619–1623, ELRA, 2010.

Suzanne Eggins and Jim R. Martin. Genres and registers of discourse. In T.A. van Dijk, editor, *Discourse as Structure and Process*, volume 1, pages 230–257, 1997.

Barbara A. Fox. *Discourse structure and anaphora: Written and conversational English*. Cambridge University Press, 1987.

Nelson Francis and Henry Kucera. *Brown Corpus Manual*. Brown University, 1979.

Michael Grabski and Manfred Stede. Bei: Intraclausal coherence relations illustrated with a German preposition. *Discourse Processes*, 41(2):195–219, 2006.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Helmut Gruber and Peter Muntigl. Generic and rhetorical structures of texts: Two sides of the same coin? *Folia Linguistica*, 39(1-2):75–113, 2005.

Michael A.K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

Michael A.K. Halliday and Christian M.I.M. Matthiessen. *An introduction to functional grammar*. Arnold, London, 2004.

Ruqaiya Hasan. Coherence and cohesive harmony. In J. Flood, editor, *Understanding reading comprehension: Cognition, language and the structure of prose*, pages 181–219. International Reading Association, Newark, 1984.

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri van der Vloet, and Jean-Luc Verschelde. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech,28-30 May 2008*, 2008.

Michael Hoey. *Patterns of lexis in text*. Oxford University Press, 1991.

Katja Jasinskaja, Jörg Mayer, Jutta Boethke, Annika Neumann, Andreas Peldszus, and Kepa Rodríguez. Discourse tagging guidelines for German radio news and newspaper commentaries. Technical report, Universität Potsdam, 2007.

Judith M.H. Kamalski. Coherence marking, comprehension and persuasion. On the processing and representation of discourse. *LOT Dissertation Series*, 158, 2007.

Andrew Kehler. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI, 2002.

Alistair Knott and Ted Sanders. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175, 1998.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse Workshop (CID08), Potsdam, Germany*, 2008.

Harald Lüngen, Csilla Puskàs, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. Discourse segmentation of German written text. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, 2006.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448, 2000.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, 1999.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

Christoph Müller and Michael Strube. MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, 2001.

Michael O'Donnell. RST-Tool: An RST analysis tool. In *Proc. of the 6th European Workshop on Natural Language Generation, Duisburg*, 1997.

Massimo Poesio, Rosemary Stevenson, Barbara D. Eugenio, and Janet Hitzeman. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004.

Livia Polanyi, Martin van den Berg, Chris Culy, Gian L. Thione, and David Ahn. A rule based approach to discourse parsing. In *Proceedings of SIGDIAL '04. Boston, MA*, 2004.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse TreeBank 2.0. Annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, 2007.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation*, 2008.

Gisela Redeker and Markus Egg. Says who? On the treatment of speech attributions in discourse structure. In *Proceedings of Constraints in Discourse II*, pages 140–146, 2006.

David Reitter. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV-Forum, GLDV Journal for Computational Linguistics and Language Technology*, 18:38–52, 2003.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Towards a taxonomy of coherence relations. *Cognitive Linguistics*, 15:1–35, 1992.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Coherence relations in a

cognitive theory of discourse representation. *Journal of Pragmatics*, 4:93–133, 1993.
Peter R. Skadhauge and Daniel Hardt. Syntactic identification of attribution in the RST treebank. In *Workshop On Linguistically Interpreted Corpora*, 2005.
Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT/NAACL 2003*, pages 228–235, 2003.
Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102, 2004.
Manfred Stede. Disambiguating rhetorical structure. *Research on Language & Computation*, 6(3): 311–332, 2008.
Michael Stubbs. Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. *The handbook of discourse analysis*, pages 54–75, 2001.
Maite Taboada. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592, 2006.
Maite Taboada and Julia Lavid. Rhetorical and thematic patterns in scheduling dialogues: A generic characterization. *Functions of Language*, 10(2):147–178, 2003.
Maite Taboada and William C. Mann. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–588, 2006a.
Maite Taboada and William C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, 2006b.
Maite Taboada, Julian Brooke, and Manfred Stede. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 62–70, 2009.
Sanna-Kaisa Tanskanen. *Collaborating towards coherence: Lexical cohesion in English discourse*. Benjamins, Amsterdam, 2006.
Gian Lorenzo Thione, Martin van der Berg, Chris Culy, and Livia Polanyi. LiveTree: An integrated workbench for discourse processing. In B. Webber and D. Byron, editors, *ACL 2004 Workshop on Discourse Annotation, Barcelona, Spain*, pages 110–117, 2004.
Sander E. J. Timmerman. Automatic recognition of structural relations in Dutch text. *MA thesis, University of Twente*, 2007.
Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80, 2009.
Thomas A. Upton. Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics*, 7(1):65–85, 2002.
Thomas A. Upton and Mary Ann Cohen. An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies*, 11(5):585–605, 2009.
Nynke van der Vliet. Syntax-based discourse segmentation of Dutch text. In Marija Slavkovik, editor, *Proceedings of the 15th Student Session, ESSLLI*, pages 203–210, 2010.
Teun A. van Dijk. *News as discourse*. Erlbaum, Hillsdale, 1988.
Bonnie Webber. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779, 2004.
Bonnie Webber. Accounting for discourse relations: constituency and dependency. In M. Dalrymple, editor, *Intelligent linguistic architectures*, pages 339–360, 2006.
Bonnie Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 674–682, 2009.
Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 29:545–587, 2003.
Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.
Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. A procedure for collecting a database of texts annotated with coherence relations. Technical report, MIT, Cambridge, MA, 2003.

On the Information Status of Antecedents: Referring Expressions Compared

Iker Zulaica-Hernández and Javier Gutiérrez-Rexach
The Ohio State University

Abstract

The differences in use among referring expressions have been explained on the basis of the information status or the cognitive status of their antecedents. Thus, for example, it has been proposed that highly accessible referents (in the current focus of attention of the discourse participants) license the use of personal pronouns while banning the use of demonstratives. This paper compares the referential properties of Spanish demonstrative expressions and the neuter personal pronoun through the study of a Spanish corpus. Our hypothesis is that these two referring expressions are very similar, if not identical, regarding the information status of their antecedents. We will argue that the difference between these expressions lies in that demonstratives actively contribute to information structure by marking topic or subtopic shifts in discourse, whereas speakers use neuter personal pronouns to refer to established topics.

1 Introduction

Over the last decades, corpus-based research has turned out to be of great importance in helping provide adequate solutions to many theoretical issues in different linguistic fields. A number of corpus studies have been conducted on the phenomenon of discourse deixis[1] achieving outstanding advances in the comprehension of the mechanisms that govern the class of referential chains that arise in discourse. Some of these studies have put the focus on providing an adequate annotation scheme for discourse deixis, other studies are focused on the quantitative part and most of them combine the two perspectives. For example, Poesio and Artstein (2008) present their annotation scheme for the ARRAU[2] corpus and tackle important questions like the referential ambiguity of certain expressions in discourse-anaphora patterns. Regarding the reliability tests on discourse deixis, these authors point out the following: "for discourse deixis we found that annotators agreed on the general textual regions that evoke the referents, though they often disagreed on the exact boundaries, resulting in agreement of around $\alpha = 0.55$" (2008:1171).

Dipper and Zinsmeister's work (2009) focuses on German and provides rigorous annotation guidelines to determine the semantic type of anaphor and antecedent. The authors justify their semantic annotation due to the idiosyncrasy of antecedents in discourse deixis, that is, the anaphoric link cannot be resolved through grammatical restrictions. Navarretta and Olsen's study (2008) is an extension of the MATE/GNOME co-reference annotation scheme (Poesio, 2004) that accounts for abstract anaphora in Danish and Italian[3]. Besides annotating the type of clausal antecedent, the semantic type of the referent (events, states, fact-like entities, etc.)

1 For an overview of discourse deixis, see the general studies by Asher (1993), Byron (2004), Fox (1987), Webber (1979) or the studies on Spanish demonstratives by Gutiérrez-Rexach and Zulaica-Hernández (2007) and Zulaica-Hernández (2008).

2 The home page of the ARRAU project is http://cswww.essex.ac.uk/Research/nle/arrau.

3 The English home page of the DAD project is http://www.cst.dk/dad.

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena", pages 173-186, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

they also annotate anaphoric distance, measured in terms of clauses in between the anaphor and the antecedent. Navarretta and Olsen draw important conclusions on the differences between Italian and Danish abstract anaphora and on how some of the proposals made by Gundel et al. (2004) concerning the relationship between antecedent and pronoun types do not hold for Italian.

Recasens (2008) conducted a corpus study of the discourse-deictic properties of Spanish and Catalan expressions, including demonstratives, based on the annotated corpus AnCora[4]. In her study, the author tests whether Webber's ideas on discourse deixis also hold for Catalan and Spanish. Besides the importance of her quantitative study, one of the most significant points in Recasen's paper concerns the intrinsic difficulty to clearly delimit the exact boundaries of the antecedents in cases of discourse deixis. Thus, she appeals to Webber's (1988, 1991) ideas on the unspecificity of the antecedent and to Poesio et al.'s (2006) theory on the underspecification of anaphora (*The Justified Sloppiness Hypothesis*) that, in essence, postulates that certain ambiguous anaphoric expressions may be left unresolved or simply not fully specified in the right context.

Other studies have placed the focus on the analysis of referring expressions and information status across languages (Bosch et al., 2003; Carminati, 2000; Kaiser and Trueswell, 2005; Kameyama, 1999; Navarretta, 2005, 2007; Sturgeon, 2008; Vieira et al., 2002). Although there is no total consensus when the referential properties of demonstratives and personal pronouns are compared, the most widely accepted thesis is that antecedents of demonstratives are most commonly non-topical whereas personal pronouns commonly have topical elements as their antecedents. Topichood is assumed to be dependent on syntactic configurations; namely, highly prominent positions (subject) are topical whereas less prominent syntactic positions (object, adjunct) are non-topical.

The aim of this paper is to compare the referential properties of Spanish demonstratives and the neuter personal pronoun *lo* ('it') as elements that participate in discourse anaphora and discourse deixis patterns. Following previous work on the information status of referring expressions (Prince, 1981b; Ariel, 1988, 1990; Gundel et al, 1993; Hegarty et al, 2003; Poesio and Modjeska, 2005), we put the main focus in checking whether there are significant differences in the referring behavior of these two Spanish linguistic expressions and whether these differences, if any, may have a bearing on the information status of their referents. With this purpose, we have conducted a corpus study where we have tested two factors that can help us distinguish these two elements, namely, the textual distance of the antecedent[5] and the morphological type of the antecedent. The corpus used in this study is the CREA corpus[6]. Contrary to what is argued by proponents of the *Accessibility Scale* (Ariel, 1988) or the *Givenness Hierarchy* (Gundel et al., 1993), our hypothesis is that Spanish demonstratives (determiners and pronouns) and the neuter personal pronoun

4 The AnCora corpora – Annotated Corpora for Catalan and Spanish (Taulé et al. 2008) – consist of two corpora of 500,000 words for Catalan (AnCora-Ca) and Spanish (AnCora-Es). The corpora are accessible from http//clic.ub.edu/ancor.

5 Referential distance has already been considered as a factor possibly influencing the degree of accessibility of different referring expressions (see Maes & Noordman 1995 and Ariel 2001 for discussion on this topic).

6 The home page of the CREA corpus is http://corpus.rae.es/creanet.html.

lo (‘it’) do not differ in their basic referring capabilities or the information status of their antecedents. Rather, we will argue that the difference between these elements lies in that speakers use demonstratives to mark topic or subtopic shifts in the discourse. This is accomplished by focusing the “hearer's attentional state” on specific discourse referents. By using this strategy, speakers would make hearers aware of a change in the general or local topic at a certain point in discourse. On the other hand, the main function of the neuter personal pronoun is to refer to topics already established in discourse or, in other words, to maintain topic continuity.

2 Information Status: Accessibility and the Current Focus of Attention

Different hierarchical scales have been proposed to account for the different distribution shown by the range of referring expressions across languages. Thus, for example, Prince (1981b) was the first to propose a hierarchy for discourse entities called the *Scale of Familiarity*, which is based on three main factors: predictability, saliency and the common knowledge shared by the speaker and addressee.

In Ariel’s (1988) *Accessibility Scale*, the notion of accessibility is defined as the relative ease with which the addressee can identify the referent of a referring expression or, alternatively, the ease with which the addressee can retrieve the intended referent from memory. According to the scale, demonstratives occupy an intermediate position in terms of the degree of accessibility they confer to their referents. As the scale clearly indicates, the less informative (null) forms (gaps, PRO, etc.) occupy the highest position, that is, they are high accessibility markers. Unstressed and cliticized pronouns also occupy a high position in the scale.

Gundel et al.’s (1993) *Givenness Hierarchy* is an implicational hierarchy of cognitive states and linguistic forms aimed to resolve the different anaphoric behavior of pronominal and non-pronominal anaphors. According to this hierarchy, the referents of demonstratives have either activated or familiar status but never in focus, whereas the referent of a neuter personal pronoun always has the status in focus. Being *activated* for a referent means that, at a given point in the discourse, there must be a representation of the referent in short-term memory. On the other hand, being *in focus* means that the referent is not only in short-term memory but also at the current center of attention. As the authors pointed out, at a given discourse point, entities *in focus* are the partially ordered subset of *activated* entities that are more likely to be the topic in subsequent discourse.

Gundel et al. (2005) analyzed the behavior of English demonstratives ‘this/that’ and the unstressed pronoun ‘it’ in the Santa Barbara corpus of spoken American English. These authors observed that demonstrative anaphors were used to refer to abstract entities in 85% of the analyzed cases, whereas only 15% of the cases were anaphorically referred to with the pronoun ‘it’. They explained this fact by assuming that material introduced in clauses (e.g. clausally introduced entities like propositions or events, which are typical antecedents for demonstrative anaphors) is *activated* compared to material introduced via noun phrases in prominent syntactic positions, which is more likely to be *in focus*.

Poesio and Modjeska (2005) tried to make the cognitive notions from the Givenness Hierarchy (i.e., *in focus*, *activated*) and short-term memory, more precise.

They primarily adopt the computational approach to anaphora resolution of *Centering Theory* (Grosz, Joshi & Weinstein, 1995) and follow previous findings on that field to better define these notions. Poesio and Modjeska annotated the corpus GNOME for this purpose and tested the following hypothesis regarding the speaker's non-preference to use This-NP's to refer to *in focus* entities:

- This-NPs are preferentially used to refer to entities other than the $CB(U_i)$, the CB of the utterance containing the This-NP.
- They are used to refer to entities other than the $CB(U_{i-1})$, the CB of the previous utterance.
- They are used to refer to entities other than $CP(U_{i-1})$, the most highly ranked entity of the previous utterance.

In Centering Theory it is assumed that new discourse entities (forward-looking centers or CFs) introduced by each utterance are ranked based on information status. The forward-looking centers of U_n only depend on the expressions that constitute that utterance; they are not constrained by features of any previous utterance in the segment. The most highly ranked entity of the forward-looking set is called the CP (the preferred center). The CB (the backward-looking center of an utterance U_n) is Centering's equivalent of the notion of topic or focus. The backward-looking center of U_i connects with one of the forward-looking centers of U_{i-1}. The $CB(U_i)$, the backward-looking center of utterance U_i, is the highest ranked element of $CF(U_{i-1})$, i.e. the CP of U_i.

The authors propose a general hypothesis regarding the speaker's preference to use This-NPs for reference to *activated* (*active* in their own terminology) discourse entities. An entity is *active* if:

- It is in the visual situation; or
- it is a CF of the previous utterance; or
- it is part of the implicit linguistic focus. They only considered as part of the implicit focus those entities that can be *constructed* out of the previous utterance. An entity can be constructed out of an utterance if: A) It is a plural object whose elements or subsets have been explicitly mentioned in that utterance; or B) It is an abstract entity introduced by that utterance. They consider two types of abstract entities:
  i. Propositions
  ii. Types[7]

7 For Poesio and Modjeska (2005), types are those cases of generic reference that have concrete objects as instances.

They got the following results in terms of distribution:

Class	Number (%)
Anaphora	45 (40%)
Visual Deixis	28 (25%)
Discourse Deixis	19 (17%)
Type	9 (8%)
Plurals	1
Ellipsis	1
Time	1
Unsure	5
Disagreement	3
Total	112

Table 1. Distribution of This-NPs (Poesio and Modjeska, 2005)

With respect to the correlation between focus and This-NPs, they found the following principal results:

- 8-11 violations to the hypothesis that a This-NP is used to refer to entities other than the CB(U_{i-1}) were found, which is therefore verified by 90%-93% of This-NPs.
- The hypothesis that This-NPs are used to refer to entities other than CP(U_{i-1}) is verified by 75-80% of This-NPs.
- The hypothesis that This-NPs are used to refer to entities other than CB(U_i) is verified by 61-65% of This-NPs.

So the hypothesis that received more empirical support is the following: This-NPs are used to refer to entities which are active but not the backward-looking center of the previous utterance. Based on these results and an in-depth study of the violation cases they proposed the version that leads to the fewest number of violations of Grice's *Maxim of Quantity* (1989):

The This-NP Hypothesis: This-NPs are used to refer to entities, which are *active* in the sense specified above. However, pronouns should be preferred to This-NPs for entities other than CB(U_{i-1}).

In a series of papers, Hegarty (2003, 2006) and Hegarty et al. (2001, 2003) studied abstract object anaphora from a semantic perspective. Generally speaking, all these studies coincide in that clausally introduced entities are more commonly referred to with a demonstrative pronoun hence indicating that the cognitive status of the entities is activated. There is an important point to be made regarding the theoretical appropriateness of the cognitive statuses as reflected in the Givenness Hierarchy. Hegarty indicates that an entity will be *in focus* only if it has been mentioned by a

nominal expression in a prominent syntactic argument position earlier in the utterance or in the previous utterance; a supposition which is compatible with *Centering Theory* and results in the experimental psycholinguistic literature. On the other hand, peripherally introduced entities, including those introduced by less prominent nominal expressions and by clauses, will be *activated* upon their introduction, placed in working memory within the field of attention, but never at the center of attention.

As we have seen so far and concerning English data, there appears to be consensus on the information and cognitive status of the entities referred to with demonstratives and the weak pronoun 'it', especially when reference to abstract entities is involved. Thus, speakers would use demonstratives to refer to *activated* entities (or *active* in Poesio and Modjeska's terminology), which rank lower than *in focus* entities regarding their cognitive and information status. Unlike demonstratives, the pronoun 'it' would be strongly preferred for reference to entities in the current focus of attention, i.e. *in-focus*. But there are reasons to believe that these findings cannot be extrapolated to all languages. For example, Navarretta (2008) found language-specific results for Danish and Italian regarding the referential behavior of demonstratives and personal pronouns. She found that the most frequently used abstract anaphor in Danish is the ambiguous *det* ('it/this/that') and her data indicate that the anaphors *det* and *dette* (demonstrative 'this') are used with all antecedent types and to make reference to all sorts of referents. Also, personal pronouns are also used in Danish with clausal antecedents. Regarding Italian, Navarretta found that zero anaphors and personal pronouns are often used in this language in contexts where demonstrative pronouns occur in English. Also, zero anaphors are the most frequently used pronouns to refer to propositions (let us remind that the referents of zero pronouns are *in focus* in the Givenness Hierarchy). These data indicate important cross-linguistic differences in the referential behavior of referring expressions and/or the information status of abstract referents. Our Spanish data appear to point in a similar direction. We present our findings in the next sections.

3 Corpus: Methodology and Results

The CREA corpus is a large linguistic database (over 160 million words) comprising several language varieties, text types and genres. Corpus queries allow users to retrieve a text fragment, situating words in context. 50% of its sources are from Spain (45 million speakers), and 50% from Latin America (350+ million speakers). 90% of the words in CREA are from written sources, and only 10% from oral sources. The CREA corpus of Spanish is not annotated so we did the annotation manually and only for the cases analyzed. The size of the corpus and the high frequency of the expressions analysed (neuter personal pronoun and demonstratives) made it unfeasible to analyze all the occurrences found.

We have analysed a total number of 327 occurrences divided as follows: 120 occurrences of the neuter personal pronoun *lo* ('it') and 207 occurrences of demonstrative expressions. All the occurrences of neuter personal pronouns analysed (n = 120) were divided into three groups corresponding to three different corpus searches: *lo entiendo* ('I understand it'), *lo necesito* ('I need it') and *lo tengo* ('I have

it'), so 40 occurrences per group were scrutinized. The reason for having analysed these particular combinations is twofold. On the one hand, this allowed us to discard other, non-referential uses of the personal pronoun in Spanish. On the other hand, these three groups would allow us to test not only referential distance but also the denotation of the antecedent and check whether it may possibly have a bearing on the cognitive status and different accessibility marking shown by the neuter personal pronoun. Thus, by using the predicates *entender* ('understand'), *necesitar* ('need') and *tener* ('have') we have tried to force different semantic readings for the antecedent. The predicate *entender* ('understand') would show a preference for higher order antecedents such as concepts, ideas or hypotheses rather than concrete, physical objects. Conversely, the verb *tener* in the expression *lo tengo* ('I have it') exhibits a preference for physical-object denoting antecedents as, under normal conditions, people have/own physical objects. The verb *necesitar* ('need') is intended to occupy an intermediate position in between the former two predicates. The aim overall was to obtain a sample ample enough to be able to draw some initial conclusions regarding the possible influence of antecedent denotation.

The first factor analysed was referential distance, that is, the distance between the anaphor and the antecedent. In order to check referential distance we segmented our examples into clauses. Our definition of a clause includes main and subordinate clauses, where the verbal phrase (VP) is taken as the clausal indicator. Thus, for example, two clauses joined with conjunction *y* ('and') count as two clauses and a main clause with a subordinate clause counts as two clauses as well. Obviously, we came across problematic cases like infinitival clauses (e.g. *Having a relationship is not in my plans for the moment*), which were also taken as a clause. The results are shown in Table 2.

Anaphor	CL_0		CL_1		CL_2		CL_3		$CL_{\geq 4}$	
	#	%	#	%	#	%	#	%	#	%
Lo necesito (I need it)	3	7.5	30	75.0	4	10.0	0	0	3	7.5
Lo entiendo (I understand it)	2	5.0	37	92.5	0	0	1	2.5	0	0
Lo tengo (I have it)	11	27.5	26	65.0	2	5.0	1	2.5	0	0
Total	16	13.33	93	77.5	6	5.0	2	1.66	3	2.5

Table 2: Referential distance for accusative personal pronoun

In total, we analyzed 207 occurrences of demonstratives (n = 207). The results of this sample are shown in Tables 3 and 4. In the first place, we retrieved a sample of 50 adnominal demonstratives divided into two groups of 25 cases each: *este hecho* ('this fact') and *ese hombre* ('that man'). The reasons for having analyzed these particular NPs are the same that we explained for the neuter personal pronoun in the previous paragraph. With the NPs *hecho* and *hombre* we analyzed different denotations of the antecedent, namely, a higher order entity and a physical entity, respectively. A second corpus search consisted of 157 cases of demonstrative pronouns: 63 instances of demonstrative pronoun *esto* ('this'), 69 of *eso* ('that') and 25 of *aquello* ('that

further'). The disparity of the analyzed occurrences of demonstrative pronouns, in particular the low number of tokens for pronoun *aquello* (25), is due to the actual frequency of use of demonstratives in modern Spanish. Overall corpus figures show that pronominal demonstrative *aquello* has a very low frequency of use (6%) compared to the frequencies shown by *esto* and *eso*. Even between these two pronouns the differences are quite relevant (*eso*: 60%) and (*esto*: 34%). Nevertheless, overall figures vary when the frequency of use as demonstrative determiners is considered. Demonstrative determiner *ese* has a frequency of 30% whereas determiner *este* shows a percentage as high as 61%. Again, demonstrative determiner *aquel* shows a rather low frequency of use (9%).

Anaphor	CL_0		CL_1		CL_2		CL_3		$CL_{\geq 4}$	
	#	%	#	%	#	%	#	%	#	%
Este hecho (this fact)	1	4.0	21	84.0	3	12.0	0	0	0	0
Este hombre (this man)	2	8.0	19	76.0	2	8.0	0	0	2	8.0
Total	3	6.0	40	80.0	5	10.0	0	0	2	4.0

Table 3: Referential distance for demonstrative determiners

Our sample of demonstrative pronouns was restricted to events as type of referents of demonstrative anaphors. In order to restrict the referential potential of demonstratives, we searched the corpus for expressions consisting of a combination of a demonstrative pronoun plus a typical predicate of events like *suceder* ('happen'), *ocurrir* ('occur') or *pasar* ('happen'); e.g. *eso sucedió* ... ('that happened...'), etc. This forces a specific denotation for the antecedent: events. The principal advantages of this strategy were to restrict the large number of demonstrative pronouns in the corpus and also eliminating potential exophoric (extra-textual) reference while having a denotation that is not particularly biased as for the morphological type of antecedent used to convey it (NP or clausal).

Anaphor	CL_0		CL_1		CL_2		CL_3		$CL_{\geq 4}$	
	#	%	#	%	#	%	#	%	#	%
Esto (this)	0	0	50	79.4	11	17.5	1	1.6	1	1.6
Eso (that)	0	0	54	78.3	5	7.2	7	10.1	3	4.3
Aquello (that further)	0	0	19	76.0	4	16.0	0	0	2	8.0
Total	0	0	123	79.0	20	12.0	8	5.0	6	4.0

Table 4: Referential distance for demonstrative pronouns

The second factor analyzed was the morphosyntactic type of the antecedent. We have included the total number of occurrences analyzed in this study (n = 327). We have considered two types: NP and Other (clausal). Within the type *Other (clausal)* we have also included infinitival clauses and other antecedents that expand beyond the clause (i.e. complex clauses or even larger text spans). The results of the study

involving antecedent type are shown in Table 5.

Anaphor	NP		Other (Clausal)	
	#	%	#	%
Lo necesito (I need it)	20	50.0	20	50.0
Lo entiendo (I understand it)	13	32.5	27	67.5
Lo tengo (I have it)	33	82.5	7	17.5
Este hecho (this fact)	6	24.0	19	76.0
Este hombre (this man)	24	96.0	1	4.0
Esto (this)	6	9.5	57	90.5
Eso (that)	16	23.2	53	76.8
Aquello (that further)	5	20.0	20	80.0
Total	123	37.6	204	62.4

Table 5: Morphological type of antecedent

In general, clausal antecedents are widely preferred (62.4%) over NP antecedents (37.6%) when all referring expressions are taken together. When we analyze the expression types individually, we found the following frequencies: the neuter pronoun *lo* shows a slight preference for non-clausal antecedents (55%) over clausal ones (45%). Some individual differences appear to be based on the type of the predicate accompanying the personal pronoun or demonstrative determiner analyzed. For example, the expression *lo entiendo* ('I understand it') shows a strong preference for clausal antecedents over NPs (27 and 13 occurrences, respectively). Conversely, the neuter pronoun in the expression *lo tengo* ('I have/posses it') shows a strong preference for NP over clausal antecedents (33 and 7 occurrences, respectively). Demonstrative pronouns (*esto*, *eso* and *aquello*) show a strong preference for clausal antecedents (90%, 76% and 80%, respectively), whereas demonstrative determiners show opposite preferences depending on the noun involved in each particular expression: the NP *este hecho* ('this fact') shows a strong preference for clausal antecedents (76%), most likely due to the denotation of the noun, whereas the NP *este hombre* ('this man') shows an even stronger preference for NP antecedents (96%).

4 Discussion

After having performed the *Chi-Square* test to check the statistical significance of our results, we can draw the following conclusions from the data presented in Tables 2–5. As far as referential distance is concerned, when we group the three categories together (i.e. personal pronoun, demonstrative determiners and demonstrative pronouns) the distribution observed is highly significant (X^2=29.999 (df = 8), p <0.0005); all three categories show a strong tendency to find their antecedents in the clause immediately preceding the anaphor (CL_1). Total frequencies are very similar for all three types of referring expressions: 77% (personal pronoun *lo*), 80%

(demonstrative determiners) and 79% (demonstrative pronouns). With only minor exceptions, a general tendency is observed that can be stated as follows: The higher the textual distance between the antecedent and the anaphor, the lower the frequency of occurrence of an antecedent-anaphor pattern.

As far as referential distance is concerned, our data show that all three anaphors show a strong preference to find their antecedent in CL1, so there appear to be no differences between demonstratives and the neuter personal pronoun in this respect. The personal pronoun *lo* though shows a somewhat significant rate of co-occurrence with antecedents in CL_0 (the clause containing the anaphor.) This is mainly due to a somewhat frequent Spanish construction that combines a demonstrative with the referential neuter personal pronoun *lo*. An example from the corpus is shown in (1).

(1) El ser humano es una bestia. Eso lo se hace años.
'The human being is a beast. I have known that since long.'

In this study, the occurrences of the neuter pronoun in this particular configuration have been included in the CL_0 group. The demonstrative pronoun *eso* refers back to the antecedent in the previous clause and the pronoun *lo*, in turn, has the demonstrative as antecedent. The relevance of this construction lies in the ability of both referring expressions to co-occur within the same clause and refer to the same discourse entity while having different morphological antecedents. Notice that the demonstrative pronoun occupies a highly prominent position within the sentence (subject), which is most commonly filled with topical elements. Also, the antecedent expression is highly salient as regards processing effort, recency of mention or memory retrieval, since it is introduced by the utterance closest to the demonstrative. In addition, the antecedent is not a subject or an object but a whole proposition. All these factors together lead us to suggest that the antecedent of the demonstrative is, contrary to expectations, a topical antecedent (e.g. either the general discourse topic or a local subtopic). Notice how the demonstrative in discourse (1) is immediately followed by the personal pronoun *lo*, which is commonly assumed to refer to highly topical entities. This co-occurrence is not obligatory, as is manifested by the ability of the pronoun *lo* to appear without the demonstrative in the same type of construction. This is shown in (2):

(2) Es algo incómodo revisar tu trabajo, pero (eso) ya lo tengo asumido.
'It is somewhat uncomfortable to revise your own work, but I have already accepted that.'

As regards the morphological type of the antecedent, we observed some highly significant distributions. When total figures for all three referring expressions are considered, we get extremely few demonstrative pronouns referring to NP antecedents (X^2=54.0238 (df = 2), p <0.0005). This is not surprising as demonstrative pronouns are most commonly used to refer to abstract entities in Spanish, so what this figure indicates is that abstract entities are most usually conveyed via clausal antecedents. Also, when we consider all demonstratives (determiners and pronouns together) and the neuter pronoun (X^2=24.4165 (df = 1), p <0.0005) we still get a very

strong preference for clausal antecedents over NPs (62.4% and 37.6%, respectively). We get similar frequencies when the neuter personal pronoun and demonstrative pronouns are compared (66.4% of clausal antecedents and 33.6% of NP antecedents, respectively) with a highly significant statistical significance (X^2=43.5814 (df = 1), p <0.0005). Although the neuter personal pronoun shows a slight preference for NP over clausal antecedents (55% and 45%, respectively), these figures are somewhat surprising, given that we did not expect to find so many cases of clausal antecedents for the neuter personal pronoun.

In view of these data, it appears that referential distance will not help us to discriminate among the referring properties of the expressions analyzed in this study. Let us recall that overall figures indicate that the preferred location of the antecedent is CL_1 for all the expressions involved. If we consider recency of mention as a factor to explain the information status of an antecedent, then we can conclude that there are no significant differences in the information status (i.e. in focus vs. activated) of the entity referred to with a neuter personal pronoun or a demonstrative expression. On the other hand, although figures indicate that demonstratives show a strong preference over the neuter personal pronoun for clausal antecedents, our data also show a high number of cases of clausal antecedents with the personal pronoun *lo* (55% and 45% of NP and clausal antecedents for the neuter personal pronoun, respectively, for n = 120).

5 Conclusions

A widely accepted thesis concerning the information status of referring expressions is that antecedents of demonstratives are most commonly non-topical whereas personal pronouns are commonly anteceded by topical elements. Topichood is commonly assumed to be dependent on syntactic configurations, that is, highly prominent positions (i.e. subject) are topical whereas less prominent syntactic positions (i.e. object or adjunct) are non-topical. When information status is defined in cognitive terms, it is commonly assumed that the referents of demonstratives occupy a lower position in terms of cognitive accessibility (activated) whereas personal pronouns mark their referents as highly accessible (in focus).

As regards Spanish in discourse anaphora/deixis uses, our main hypothesis is that demonstratives and the neuter personal pronoun do not differ much in the way they refer to discourse entities. We have studied two factors in a corpus of Spanish, which are directly related to the referential properties of these elements: distance of the antecedent and morphological type of the antecedent. The data indicate that antecedent distance is not a distinguishing factor as all the expressions analysed showed a strong preference to find their antecedents in the clause that is the closest to the anaphor. Thus, if we consider antecedent distance as a factor having an effect on the information status of the antecedent (i.e. most recent antecedents are more accessible than antecedents located at a greater distance), then demonstratives and the neuter personal pronoun show a very similar behavior. On the other hand, the resulting figures show that demonstrative pronouns have a clear preference for clausal antecedents over NP ones but we also found a significant number of cases of the neuter personal pronouns with clausal antecedents. This may be due to the

denotation of the referent involved, since the referents of clausal antecedents are most commonly abstract entities such as propositions, facts, events, etc.
The main empirical conclusions can be summarized as follows:

- Demonstratives and the neuter personal pronoun alike show a strong preference to find their antecedents in the clause closest to the anaphor (CL_1).
- Demonstratives and the neuter personal pronoun alike can refer to abstract entities (propositions, facts, etc.)
- Demonstratives show a stronger preference for clausal antecedents but the neuter personal pronoun shows a high number of clausal antecedents (45% out of total number of cases analysed of the neuter personal pronoun.)

In many respects, the results from our study coincide with previous cross-linguistic research in the sense that different languages show language-specific characteristics in the way they realize abstract pronominal anaphora. In particular, our data resemble the findings by Fraurud (1992) who found no differences between the Swedish anaphor *det* (ambiguous 'it/that/this') and the demonstrative anaphor *detta* ('this') in abstract reference; or the findings by Navarretta (2008) on Danish and Italian mentioned earlier. In our view, our data appear to confirm our initial hypothesis that the main role of Spanish demonstratives in discourse-anaphora/deixis uses involves marking (sub)-topic shifts in discourse. This procedure should be conceived as an instruction on the part of the speaker for the addressee to focus on a particular discourse entity and with a precise communicative intention, i.e., making her interlocutor aware that a topic shift is taking place or a new local subtopic has been introduced. In terms of cognitive or information status, demonstratives are devices used by speakers to bring entities into the current focus of attention. We defend that this focusing property closely resembles that of demonstratives in deixis proper (i.e. use of demonstratives to point to physical entities) or nuclear pitch accent in phonological focus marking. Additional support in favor of our hypothesis comes from a Spanish specific construction consisting of a demonstrative anaphor and a neuter personal pronoun both co-occurring within the same clause, next to one another and co-referential *eso lo* ('that it'); see examples (1) and (2).

References

Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(1): 65-87, 1988.

Mira Ariel. *Accessing Noun Phrase Antecedents*. Routledge, London/New York, 1990.

Mira Ariel. Accessibility theory: an overview. In T. Sanders, J. Schilperoord and W. Spooren, *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam: John Benjamins, pages 29-89, 2001.

Nicholas Asher. *Reference to Abstract Objects in Discourse*. Dortrecht: Kluwer, 1993.

Peter Bosch, Tom Rozario, Yufan Zhao. Demonstrative pronouns and personal pronouns. German *der* vs. *er*. *Proceedings of EACL2003, Workshop on the Computational Treatment of Anaphora*, 2003.

Donna Byron. Resolving pronominal reference to abstract entities. Technical report 815, University of Rochester, 2004.

Maria Nella Carminati. *The Processing of Italian Subject Pronouns*. PhD thesis. University of Massachussets, 2002.

Stefanie Dipper and Heike Zinsmeister. Annotating discourse anaphora. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP 2009, pages 166-169, 2009.

Barbara A. Fox. *Discourse Structure and Anaphora*. Cambridge, UK: Cambridge University Press, 1987.

Kari Fraurud. *Processing noun phrases in natural discourse*. PhD Dissertation, Stockholm University, 1992.

Herbert P. Grice. Presupposition and conversational implicature. In H. P. Grice (ed), *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press, pages 269-283, 1989.

Barbara J. Grosz, Scott Weinstein and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2): 203-225, 1995.

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language* 69: 274-307, 1993.

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. Demonstrative pronouns in natural discourse. In A. Branco, T. McEnery and R. Mitkov (eds.), *Proceedings of DAARC 2004*, pages 81-86, 2004.

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. Pronouns without NP antecedents: how do we know when a pronoun is referential? In A. Branco, T. McEnery & R. Mitkov (eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*. Amsterdam: John Benjamins, pages 351-365, 2005.

Javier Gutiérrez-Rexach and Iker Zulaica-Hernández. Abstract reference and neuter demonstratives in Spanish. In *Proceedings of DAARC 2007*, pages 25-30, 2007.

Michael Hegarty, Jeanette K. Gundel and Kaja Borthen. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27: 163-186, 2001.

Michael Hegarty. Semantic types of abstract entities. *Lingua*, 113: 891-927, 2003.

Michael Hegarty, Jeanette K. Gundel and Kaja Borthen. Cognitive status, information structure and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12 (3): 281-299, 2003.

Michael Hegarty. Type shifting of entities in discourse. In K. von Heusinger and K. Turner (eds.), *Where Semantics meets Pragmatics, Current Research in the Semantics/Pragmatics Interface*, 16. Amsterdam, Elsevier, pages 111-128, 2006.

Elsi Kaiser and John C. Trueswell. Investigating the interpretation of pronouns and demonstratives in Finnish: going beyond salience. In E. Gibson and N. Pearlmutter (eds.), *The Processing and Acquisition of Reference.* Cambridge, MA, MIT Press, 2005.

Megumi Kameyama. Stressed and unstressed pronouns: complementary preferences. In P. Bosch and R. van der Sandt (eds.), *Focus, Linguistic, Cognitive and Computational Perspectives*. Cambridge: Cambridge University Press, pages 306-321, 1999.

Alfons A. Maes and Leo G. M. Noordman. Demonstrative nominal anaphors: a case of nonidentificational markedness. *Linguistics,* 33: 255-282, 1995.

Costanza Navarretta. Combining information structure and centering-based models of salience for resolving Danish intersentential pronominal anaphora. In A. Branco, T. McEnery and R. Mitkov (eds.) *Anaphora Processing. Linguistic, Cognitive and Computational Modeling*. Amsterdam: John Benjamins, pages 329-350, 2005.

Costanza Navarretta. A contrastive analysis of abstract anaphora in Danish, English and Italian. In A. Branco, T. McEnery, R. Mitkov and F. Silva (eds.) *Proceedings of DAARC 2007 - 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 103-109, 2007.

Costanza Navarretta. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In C. Johansson (Ed.), *Proceedings of the Second Workshop on Anaphora Resolution*, pages 63-71, 2008.

Costanza Navarretta and Sussi Olsen. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*, pages 2046-2052, 2008.

Massimo Poesio. The MATE/GNOME proposals for anaphoric annotation, revisited. In

Proceedings of the 5th SIGDIAL Workshop, pages 154-162, 2004.
Massimo Poesio and Natalia N. Modjeska. Focus, activation and this-noun phrases. In A. Branco, T. McEnery and R. Mitkov (eds.), *Anaphora Processing*. John Benjamins, pages 429-442, 2005.
Massimo Poesio, Patrick Sturt, Ron Artstein and Ruth Filik. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42: 157-175.
Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*, pages 1170-1174, 2008.
Ellen F. Prince. Toward a taxonomy of given-new information. In P. Cole (Ed), *Radical Pragmatics*. New York: Academic Press, pages 223-256, 1981.
Marta Recasens. Discourse deixis and coreference: evidence from AnCora. In C. Johansson (Ed), *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*. NEALT Proceedings Series Vol. 2, pages 73-82, 2008.
Anne Sturgeon. Topic and demonstrative pronouns in Czech. In G. Zybatow, L. Szuchsich, U. Junghans and R. Meyer (eds.), *Formal description of Slavic languages*. Berlin: Peter Lang, 2008.
Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang and Gabriel Othero. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium*, pages 233-238, 2002.
Bonnie L. Webber. *A Formal Approach to Discourse Anaphora*. Garland, New York, 1979.
Bonnie L.Webber. Discourse deixis: reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113-122, 1988.
Bonnie L.Webber. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2): 107-135, 1991.
Iker Zulaica-Hernández. *Demonstrative pronouns in Spanish: a discourse based study*. PhD dissertation, The Ohio State University, 2008.