Corpus Linguistics and Information Structure Research

Anke Lüdeling, Julia Ritz, Manfred Stede, Amir Zeldes

1. Introduction

The recognition that certain aspects of sentences or utterances depend on previous discourse, speakers' knowledge management and packaging strategies, and are not fully describable in a narrow syntactic formalism has led to the formulation of Information Structure (IS) models.[1] IS models differ in the number of concepts, phenomena, and theoretical assumptions they make but all share the idea that abstract IS notions lead to surface effects, such as word order (see Chapters 21-22), accent placement (Chapters 10-12) or definiteness effects. Ideally, IS models should explain and predict the form an utterance will take, given an IS setting. Effects of IS must therefore be observable (and IS models testable) using authentic utterances. This is why corpora are interesting for studying IS theories.

In most papers, however, IS is treated in the context of isolated, not necessarily naturally occurring examples, where categories such as topicality, focus and information status are more or less well understood. For example, in (1)-(3), although there are different views of the internal structures and mechanisms marking IS, and even of the exact underlying categories, it is fair to say there is general agreement at least regarding the heads of the phrases corresponding to topic and comment (1), focused constituents (2) and givenness (3):[2]

(1) (What did he do then?) $[\text{He}]_{\text{TOPIC}}$ $[\text{bought some flowers}]_{\text{COMMENT}}$

(2) (Which one of you is Ted?) That's $[\text{ME}]_{\text{FOCUS}}$

[1] Some models locate IS in syntax, some in pragmatics, and some view it as an interface phenomenon (see the chapters in Part I). These differences will not play a central role in this chapter and we will therefore remain agnostic on this issue.

[2] In this paper we use the concepts and definitions from Krifka (2008).

(3) (A woman was walking down the street with her daughter.) Suddenly [the woman]$_{GIVEN}$ saw [a big Labrador]$_{NEW}$

Examples like these have proven valuable in understanding the motivation behind IS marking and in evaluating the coverage of different proposals in the literature. All concepts of IS must, however, – after the theory has been formulated – at some point be evaluated using empirical data. One type of data, in addition to experimental data (see Chapters 28 and 31) or questionnaire data (e.g. using QUIS: QUestionnaire on Information Structure, Skopeteas et al. 2006) – is corpus data. We use the term *corpus* to refer to naturally occurring spoken/written data collected for a linguistic research question (for general information on corpus linguistics see Lüdeling & Kytö 2008/2009). Compared to the simpler question and answer context which is dominant in the literature, there is a considerable jump in complexity for resolving and agreeing on IS categories in arbitrary, naturally occurring data such as (4)-(6).[3]

(4) (Teenagers in Zossen want a music café.) [They]$_{TOPIC?}$ [demanded]$_{TOPIC?}$ [this]$_{TOPIC?}$ in the first Zossen Roundtable on Tuesday evening.

(5) (Later I got to know Heidi.) The weeks [began]$_{FOCUS?}$ in which I would login to chat forums as Paulus.

(6) (It's unclear what interest in peace the guerillas might have.) [The logic of violence]$_{GIVEN/NEW?}$ is pervasive.

The difficulties in interpreting such examples come from many sources, including complex assumptions about common ground that are hard to verify in naturalistic settings and the absence of pragmatic context and phonetic detail (particularly intonation and stress) in interpreting written language, but also the complex conceptual structure that underlies

[3] Example (4) is translated from the Potsdam Commentary Corpus , Examples (5) and (6) are translated from TüBa-D/Z (see Section 32.3.2).

language. In (4), it is clear that the new information is the spatiotemporal setting for the previous proposition ("in the first Zossen Roundtable on Tuesday evening"). But what is the topic? Are we being told something about the teenagers "they"? The music café "this"? Or about the demand itself? In the latter case, the topic seems to be packed into a finite verb, which is not a possible locus for topicality in most approaches. In (5) the main clause informs us that 'some weeks began', but if we take, for example, an Alternative Semantics approach (Rooth 1992, Krifka 2008, and Chapter 2), it is unclear what is being focused and what alternatives are in play. It is much more the subordinate clause that gives room for alternatives (not logging in at all; not as Paulus; etc.); another solution is that the focus may correspond to the entire proposition. Finally in (6), a discourse about guerilla warfare ends with a statement about the definite subject "the logic of violence". This referent has not been mentioned, nor does it form a natural bridging relationship with previous referents (while violence is perhaps implicit in guerilla warfare, its 'logic' is not an expected part). Yet it is realized as definite, suggesting that this conclusion is somehow apparent from the previous discourse. The examples illustrate that it is not always straightforward to determine IS categories in authentic data because the context is more complex and the IS settings cannot be controlled. In this chapter we want to show that authentic examples nevertheless provide interesting insights into IS theory.

This chapter deals with both possibilities and problems of using corpus data for IS studies. Our focus is on methodology rather than specific results. Corpora can be used to explore a phenomenon (qualitatively or quantitatively) or to test hypotheses derived from a theory or model. In all cases the primary data must be *interpreted:* We must decide whether a stretch of text (a word, phrase, sentence, or some other unit) belongs to a given category or not. One of the advantages of using corpora lies in making the interpretation explicit by *annotating* data. In Section 32.2 we discuss the design and composition of IS corpora in terms of text types, data collection settings, and corpus architecture. Section 32.3 deals with

annotation of IS categories and gives an overview of existing IS corpora. In particular we will describe the evaluation and reliability of IS and coreference annotation. Section 32.4 describes the qualitative and quantitative evaluation of IS phenomena using direct and indirect evidence, and Section 32.5 draws conclusions.

2. Corpus design

Corpus design, i.e. how much of what kinds of texts are included, determines to a certain extent how a corpus can be used, especially if one wants to make quantitative statements. But even is a corpus is used merely as an 'example bank', its design may be relevant because given structures and contexts will only be found in certain corpus types.

IS phenomena depend on many linguistic and extra-linguistic variables. While the linguistic variables have been studied often – usually in controlled settings using questionnaires or introspection – extra-linguistic variables like text type, mode of communication, socio-economic status, etc. are not yet well understood.[4] This leads to the (strange) situation that sometimes results from studies that have been carried out on a relatively small and restricted corpus are generalized in ways that are probably not permissible (see Gries 2006 and Kilgarriff 2012 for discussion). In this section we focus on the influence of extra-linguistic variables on linguistic variation and consequences for corpus design.

2.1 Task-based corpora

Because evaluating IS requires good understanding of mental constructs such as common ground, contrast, or salience, an underlying task or elicitation context is often helpful in limiting the range of possible utterances, e.g., to a set of known referents whose discourse

[4] Socio-linguistic research and variation research over the past 60 or so years has revealed many factors influencing the choice of one variant over another when expressing a linguistic variable. The earliest studies focused on socio-economic variables for speakers (Labov 1972), and later studies looked at group-hierarchical relations between speakers. Other factors found to play a role are text type, mode, register (Biber 2009), whether a text is translated, or even biological relationships (Golcher 2012).

status can be controlled. This is also why question-answer settings are useful in eliciting IS data, especially in languages for which the full range of possibilities is not yet understood by researchers. Data from such settings is typically collected using questionnaires like QUIS. Although questionnaire data is not 'naturally-occurring', it nevertheless incorporates natural variation, as respondents can vary while providing similar answers. Questionnaire data makes it possible to study relationships between different forms of expression and the IS setting constructed by the task. A main advantage is that IS factors can be more easily controlled for and varied (referents can be introduced in different orders, different versions of each question can be given to each respondent). The disadvantage is, however, that results may not generalize well to unrestricted language use: As mentioned in the previous section, most utterances in unrestricted language are not answers to questions explicating topical or focal information. Just how difficult or reliable the annotation of such categories is for a given corpus will be discussed in Section 32.3.2.

Beyond questionnaire data, it is possible to collect corpora in task-based settings that allow for more than question-answer pairs, the advantage being that responses from multiple trials are comparable across respondents and languages, and the scope of reference is limited, but linguistic behavior is less constrained. A well-known type is Map Task corpora (cf. Anderson et al. 1991) where the specific map and collaborative nature of the task restrict the linguistic possibilities somewhat, yet dialogue develops in a more 'natural' way (especially interesting are situations where hearers indicate that they could not follow instructions – this can point to an IS problem, such as misunderstanding which of a set of alternatives is meant; a Map Task is included in QUIS). Another task-based corpus which has been used for IS studies is the longitudinal learner corpus of reading comprehension CREG (Meurers et al. 2010, Ott et al. 2012) which collects answers to questions about texts.

Finally, it is possible to design 'virtual corpora' by sampling e.g. only question-answer pairs from a larger corpus of spontaneous language. In this case the corpus designer can

ensure variability by including any number of constructions. For example, the German Multiple Fronting Corpus (Bildhauer 2011) contains only sentences with multiple fronted constituents and their immediate contexts. Other virtual corpora constructed for IS purposes include a corpus of OVS sentences from German newspapers (Weskott et al. 2011) or a multilingual collection of cleft sentences (Bouma et al. 2010). Such corpora are useful for studies of specific phenomena, but quantitative statements may become problematic and care must be taken that data filtering does not lead to conclusions that would be refuted by the data that has been left out.

2.2 General-purpose corpora

Many corpora used in IS research are not constructed specifically for IS. One example is SWITCHBOARD, a collection of phone conversations in American English (cf. Section 32.3.2). Often even general 'reference' corpora[5] are used. The advantage in working with less restricted corpora is that results can be generalized more easily: What applies to large, heterogeneous masses of texts is likely to lead to more robust findings. However, getting appropriate data from such corpora is not always easy. Reference corpora typically focus on written data, which contains some discourse phenomena only rarely and is missing prosodic marking which can be critical in interpreting prominence in general and IS categories in particular. Absence of prosodic marking is one of the difficulties in reliably annotating IS. For this reason, multimodal spoken corpora are particularly valuable for IS studies.

In some cases, however, only written data is available, or it is impractical to collect sufficient amounts of spoken data. The first case applies especially in historical studies of languages for which we have no spoken records (cf. Chapter 27 on Historical Linguistics). In older texts, many other composition problems infringe on the designer's freedom, such as missing or non-comparable genres across time, but also limitation to translated texts which

[5] Reference corpora are typically relatively large corpora with a principled design. The terms 'balanced' (for text types) or 'representative' are also sometimes used but typically not in a statistically sound way, see Biber (1993) and Kilgarriff (2012).

may contain translation effects that make evaluating IS categories difficult. For example, the Tatian Corpus Of Deviating EXamples (T-CODEX, Petrova et al. 2009) contains those sentences from an Old High German translation of a Latin text where word order is not identical to the original, on the assumption that these reflect independent Germanic word orders and IS phenomena.

The second case of limited accessibility applies for less studied languages, for which quantitative work cannot realistically rely on field recordings. If such languages have printed media or Internet resources, as is the case for many less studied African languages, then a written corpus may be used to supplement field work in order to get an overview of possible competing constructions and some quantitative results on their distribution and markedness (cf. Chiarcos et al. 2011 for some IS annotated resources for African languages).

In sum, the advantage of corpora of naturally occurring language is that they offer the most 'authentic' picture of language use. Ideally, only the research question should determine which corpus is chosen and researchers should know the corpus well enough to understand how to interpret the data and where extrapolation to unseen data is possible (e.g. using statistical models). The advantage of task-based corpora or corpora collected with specific IS studies in mind is that parameters can be set, controlled, and varied based on researchers' needs, reducing available alternatives. The latter factor is especially important if annotation of IS is planned, which is costly and difficult, but can make great contributions to theoretical models by formalizing an explicit analysis of the data. We therefore turn next to the prerequisites and problems concerning IS annotation.

3. Corpus Annotation

Annotation is an explicit categorization of data according to a predetermined, ideally deterministically decidable scheme (see Lüdeling 2011 for more on annotation). Most straightforward is the use of annotations as an index for finding all cases of a certain category.

But the real potential of annotations in offering new insight is when they are combined to find environments that correspond to relevant phenomena in forms not previously thought of. We therefore begin by discussing corpus architectures enabling such studies. We then give an overview of IS-annotated corpora (Section 32.3.2) and their evaluation (Section 32.3.3). We continue with comments on the possibility of using non-IS annotations for the study of IS, with an example from thetic sentences in German (Section 32.3.4) and conclude with a discussion of automatic IS annotation (Section 32.3.5).

3.1 Corpus architecture

IS is by nature a multidimensional phenomenon. There are different axes along topic/comment distinctions, focus/contrast marking, and different degrees of givenness or information status. These theoretical notions influence surface phenomena on many levels, such as phonology, morphology, syntax, semantics and pragmatics, meaning that corpora geared toward IS studies are often richly annotated with not only IS categories, but also many further annotations distinct from IS *per se*, such as parts-of-speech, syntax trees, prosodic annotation or coreference. As a result, IS corpora benefit greatly from so-called multilayer architectures, which permit the representation of arbitrarily many independent layers of annotation. On the technical side, this often means choosing stand-off XML formats such as NXT or PAULA XML (see Carletta et al. 2003 and Dipper 2005 respectively), where each annotation layer can be represented in a separate file without interfering with other annotations, which allows for the retroactive update, addition or deletion of annotation layers as research progresses.

The representation chosen should also for evaluating interdependencies between annotation layers, as well as consistency checking for potentially unreliable annotation layers (see Section 32.3.3). Zeldes et al. (2009) describe ANNIS (ANNotation of Information Structure), an open-source browser-based search and visualization platform for multilayer

corpora which has been used extensively for IS corpus studies. Figure 1 illustrates annotations in a German corpus (PCC, Stede 2004) using ANNIS.[6] By representing syntax trees, IS and coreference, among many other annotations, interactions between these levels of analysis can be studied (see Section 32.3.4).

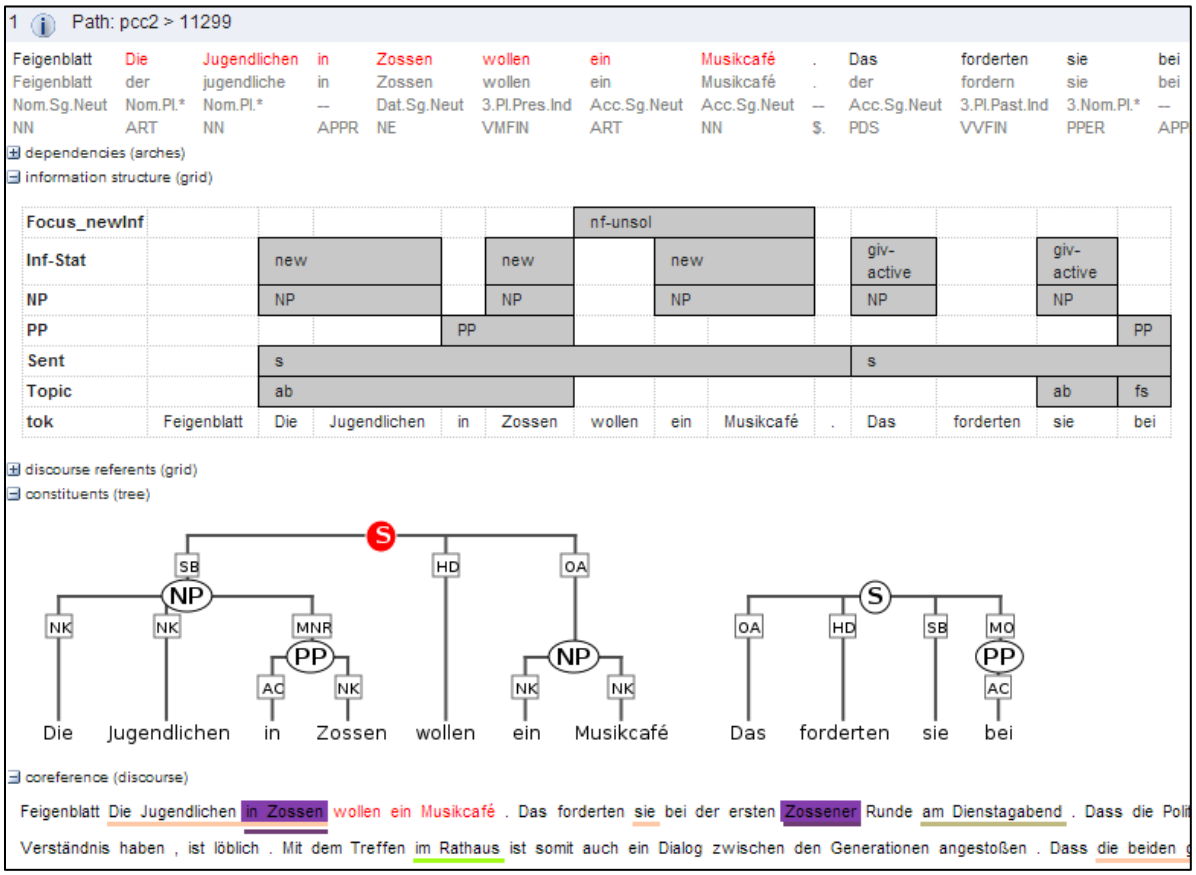


Figure 1. Multilayer annotations of syntax, IS and coreference for the sentence "Teenagers in Zossen want a music café" in ANNIS.

3.2 IS annotated corpora

While there IS-related annotations are diverse,[7] we can distinguish two major types: IS annotations proper provide categories for notions like information status, topicality or focus, which are marked directly for relevant words and phrases. Other annotations relate to IS

[6] IS and coreference annotations were done using EXMARaLDA (Schmidt & Wörner 2009) and MMAX2 (Müller & Strube 2006) respectively.

[7] This section does not aim at a comprehensive overview of IS-annotated corpora but rather wants to illustrate the breadth of IS categories that are used as well as the different types of corpora and languages used in IS studies.

phenomena indirectly, for example in the form of NP form, definiteness annotation or coreference annotation. The latter implies discourse-givenness, a sub-phenomenon of information status. Another annotation layer that is not IS proper but marks a surface phenomenon often assumed to be influenced by IS is prosody. Finally, (some) word order effects can only be found if syntax is annotated.

A starting point for corpora containing coreference and related annotations (e.g. definiteness, NP form) is represented by the MUC-6 und 7 corpora (Chinchor 2001, Chinchor & Sundheim 2003). These are English newswire corpora, though the annotation scheme has also been applied to Dutch (Hendrickx et al. 2008). Some criticism of the scheme can be found in van Deemter and Kibble (2000), which has also been addressed in the more recent OntoNotes corpus scheme (Weischedel et al. 2007), applied to English, Chinese and Arabic newswire and broadcast news. OntoNotes contains the Wall Street Journal Section of the Penn Treebank (Marcus et al. 1993) and thus contains syntactic constituent trees and coreference annotation.

In the MULI project (Baumann et al. 2004), corpus sections in English (PennTreebank, Marcus et al. 1993) and German (TIGER, Brants et al. 2002) were annotated at the levels of prosody, syntax, and discourse semantics (including coreference and bridging links) for the purpose of IS studies. TüBa-D/Z (Tübinger Baumbank Deutsch/Zeitungssprache, 'Tübingen Treebank of German Newswire', Telljohann et al. 2012; Naumann 2006) is the most extensive corpus with coreference annotation in German. It contains newspaper texts with part-of-speech and lemmatization, constituent syntax annotation, grammatical function, topological fields, and coreference. The annotation scheme distinguishes six coreference relations (coreferential, anaphoric, bound, cataphoric, split antecedent and instance), and requires explicit marking of expletives (see Section 32.3.4 for a study using this corpus).

Among the corpora containing direct annotations of IS notions, annotation of information status is most widespread. DIRNDL (Discourse Information Radio News Database for Linguistic analysis, Eckart et al. 2012) is a corpus of German broadcast news. The data is automatically annotated with LFG syntax analyses. A subset is manually annotated with prosodic information and information status (Riester et al. 2007). The annotation scheme is organized hierarchically, with the top level categories /old/, /mediated/ and /new/.

In some cases, information status annotation is complemented by partial coreference annotation. For example, Nissim et al. (2004) and Nissim (2006) describe information status annotations for the SWITCHBOARD corpus of American English telephone conversations (Godfrey et al. 1992). Their scheme is hierarchically structured and distinguishes 'old' (given), 'mediated' (comparable to the category 'accessible') and 'new' expressions. The categories 'old' and 'mediated' have 6 and 9 subcategories respectively, and include linking to the antecedent where applicable.

Annotation of topic and focus is generally more difficult (see next section) and less commonly found. The Prague Dependency Treebank (Hajič et al. 2006) is a corpus of Czech annotated with lemmas, part-of-speech tags and morphological categories, dependency syntax, coreference, and Prague school topic-focus articulation with its categories 'contextually bound', 'contrastively contextually bound', and 'contextually non-bound'. The Danish corpus DanPass was annotated with topic/focus in a study by Paggio (2006). The corpus consists of spoken monologues and dialogues (including map task dialogues) and includes phonetic and orthographic transcriptions and POS tags.

The German Multiple Fronting Corpus (Bildhauer 2011) is a corpus of sentences with more than one constituent in the prefield (the first phrase before the verb in German main clauses). Sentences were extracted from corpora available at the Institut für Deutsche Sprache (IDS). Each sentence includes its immediately preceding and following context if available.

Annotations extend to lemmas and parts-of-speech and target sentences are manually annotated with topological fields. The (multiple) fronted preverbal constituents of target sentences have been manually annotated with syntactic category and function, as well as topic, focus and givenness according to Dipper et al. (2007).

For non-European languages, Chiarcos et al. (2011) give an overview of corpus resources created at SFB632 from a typological perspective. Data was elicited with QUIS and includes audio files and prosodic annotation. The materials originate, among others, from Chadic, Gur and Kwa languages. The data is manually transcribed and glossed, and annotated for IS according to Dipper et al. (2007). There has also been work on multilingual or parallel annotation of information structure, notably Komagata (1999), Johansson (2001).[8]

Finally there are corpora which combine IS and coreference annotations. A prominent example is ARRAU (Poesio & Artstein 2008), which is based on GNOME (Poesio 2004a), MATE (Poesio et al. 1999; Poesio 2004b), and the so-called Vieira-Poesio corpus (Poesio & Vieira 1998). ARRAU comprises written and spoken English data, again including Wall Street Journal texts from the Penn Treebank and SWITCHBOARD dialogues (Godfrey et al. 1992). Data is annotated with syntactic constituents (from the Penn Treebank or manually corrected parser output), morphological features (number, gender, person), grammatical function, animacy, concrete/abstract distinction, referentiality (anaphoric, discourse-new or non-referential), as well as coreference and bridging.

The Potsdam Commentary Corpus (PCC, Stede 2004) consists of German newspaper commentaries and contains a wide variety of annotation layers for IS and more, including part-of-speech tagging, lemmatization, morphological annotation, constituent syntax as well as dependencies for a small subset of the corpus, rhetorical structure based on Rhetorical Structure Theory (RST, Mann and Thompson 1988), discourse referent annotations and

[8] We thank an anonymous reviewer for pointing these out. For parallel historical corpora see PROIEL below.

coreference. The corpus contains IS annotations according to Dipper et al. (2007), including focus marking, topicality (aboutness and frame-setters), as well as information status.

There are also some historical corpora combining information-structural and related annotations (see also Chapter 27 in this volume). They are particularly challenging to annotate because of the lack of prosodic information and uncertainty whether word orders reflect the language stage in question or whether they are influenced by translation effects or poetic constraints. The Tatian Corpus Of Deviating EXamples (T-CODEX, Petrova et al. 2009), for example, is based on the translation of Tatian's Gospel harmony from Latin into Old High German (OHG). Targeted at the study of OHG syntax, it contains only the examples where OHG word order differs from the Latin original. The corpus is annotated with parts-of-speech, syntactic categories and grammatical function, syllable count for each constituent, as well as clause status (main and subordinate clause subtypes, e.g. 'causal' etc.). The corpus also contains IS annotations according to Dipper et al. (2007) for information status, topic and focus.

Haug et al. (2009) present the parallel PROIEL corpus of the Greek New Testament and its translations into Gothic, Latin, Old Church Slavonic and Armenian. It is manually annotated with morphology and dependency syntax, animacy and parallel alignment. IS annotations include topic, givenness, and anaphoric coreference. Haug et al. (to appear) discuss how markables (the annotated units; in this case mostly DRT discourse referents) are chosen and annotated.

The overview above shows that there has been considerable activity in this field –IS-annotated corpora exist for many languages, including historical corpora. Almost all corpora with IS annotations also include other annotation layers. The main IS category annotated is information status and the main related category is coreference. Explicit annotation of topic and focus often follows the guidelines in Dipper et al. (2007). Major differences between annotation schemes for coreference and information status include categorization of deictic

pronouns, pronouns referring back to generic nominals (*their* in (7) below would receive the tag ‘coreferential’ in OntoNotes, ‘bound’ in TüBa-D/Z, or ‘old/ident_generic’ in Nissim’s scheme), as well as event anaphors and anaphors aggregating referents mentioned in the previous context (*they* in (8) receives ‘no annotation (for technical reasons)’ in OntoNotes, ‘multiple phrases’ in ARRAU, or ‘split_antecedent’ in TüBa-D/Z).

(7) [Parents] should be involved with [their] children’s education at home, not in school.

(8) [Sam] met [Pat] on the Royal Mile. [They] had a coffee and a chat, and then went back to work.

Schemes also differ regarding which semantic relations other than coreference they allow for, e.g. possession, part-of, belonging to the prototypical situation etc. For a detailed comparison of publicly available corpora and annotation schemes, see Ritz (to appear).

3.3 Evaluation of IS annotation

All annotation is interpretation. This means that it is necessary for each annotation layer to specify the tag-set (possible categories) and guidelines detailing the exponent (a word, span of words, sentence, paragraph, etc.) and assignment rules for each category. To understand the decisions that underlie an annotation layer it is crucial that guidelines are publically available (existing IS guidelines include Hajičová et al. 2000, Nissim et al. 2004, Paggio 2006, Dipper et al. 2007, Riester et al. 2010, see also Haug et al., to appear). Annotation can be done manually or automatically, though consistency must be evaluated in both cases. In this section we focus on the evaluation of manual IS annotation where the tag-set and guidelines are given to human annotators. There are several measures for evaluating inter-annotator agreement (IAA).[9] The studies below show that IS categories such as information status, focus and topic

[9] Other terms are inter-rater reliability or inter-coder consistency. Many of the measures have been used in the social sciences for a long time. Cohen’s kappa, which measures agreement between two annotators has been introduced to corpus linguistics by Carletta (1996), and kappa-values above .8 are generally considered

are extremely difficult to annotate consistently. Some of the difficulties are due to unclear tag-sets or guidelines. These could, in principle, be remedied by refining the guidelines. More interesting for theoretical reasons are those annotation problems that point to genuine problems in IS models.

The IS-related category that is easiest to annotate consistently is coreference.[10] The results for genuine IS categories are somewhat less positive. Information status,[11] which is the most widely annotated category, is theoretically interesting because it is an inherently gradual category (Krifka 2008) that is typically mapped onto a discrete number of tags. While 'given' and 'new' (under whatever label) are typically clearer, there is often disagreement concerning intermediate categories ('mediated', 'accessible'), where semantic relations need to be categorized (part-of relation, possession, is-part-of-situation). Results reported for topic and focus annotation are mixed, but typically not very good.[12]

The only larger-scale IS evaluations we are aware of are based on the guidelines in Dipper et al. (2007). Ritz et al. (2008) evaluate agreement on information status, topic and focus across different text types: question/answer pairs, map task dialogues from QUIS, and

---

satisfactory; but see Artstein & Poesio (2008) and Powers (2012) for criticism and other measures. Automatic annotation is often evaluated against a manually constructed gold standard.

[10] For MUC's coreference annotation, agreement is reported to be "in the low 80's for precision and recall", with 16% of disagreements reflecting genuine disagreement about the correct annotation (Hirschman et al. 1998). In OntoNotes, "average agreement scores between each annotator and the adjudicated results were 91.8% [for coreference]", according to Hovy et al. (2006). ARRAU has been evaluated using Krippendorff's (1980) α: Poesio & Artstein (2005) and Poesio & Artstein (2008) report agreements "in the range of $\alpha \approx 0.6–0.7$" for multiple (up to 20) annotators. For German, IAA for TüBa-D/Z achieved 83–85% f-measure (Versley 2006).

[11] For Nissim's (2006) annotations of SWITCHBOARD, the following values are reported: $\kappa = .788$ for a fine-grained hierarchical tag-set with 16 categories, and $\kappa = .902$ for old vs. mediated/new (for two annotators). The scheme makes the task somewhat easier by defining that no annotation is applied to temporal, spatial and numeric expressions. Instances annotated as 'non-applicable' (non-referring) or 'not understood' by at least one annotator were also excluded from the evaluation.

Riester et al (2010) evaluate the DIRNDL annotation scheme (Riester et al. 2007) with two annotators, reaching $\kappa = .66$ for 21 categories, and .78 for a core variant using six categories: GIVEN, SITUATIVE, BRIDGING, ACCESSIBLE, INDEF, and OTHER.

Hempelmann et al. (2005) report $\kappa = .72$ for a six-category distinction of information status based on Prince's (1981) annotation scheme applied to fourth grade textbook narratives. They specify different intuitions on semantic relations (prototypical situative knowledge) as the main source of disagreement between annotators (the categories 'contextually inferable' vs. 'new').

[12] Paggio (2006) reports on the DanPass corpus of spoken Danish. Topic and focus are defined to be disjoint. κ values, calculated from two annotators on two sections of the corpus are 0.7–0.8. The main differences she reports are whether auxiliaries form part of the focus domain or not. This results from unclear guidelines (which should specify the possible/maximal exponent for each category).

newspaper commentaries from PCC. The results for information status range from kappa values of 0.6–0.8 for top-level distinctions (given/new/accessible) and lower values of 0.55–0.73 for a finer seven-way distinction. Results for topic range 0.46–0.91, and for focus between 0.41–0.62, both considerably worse than information status. The highest values were achieved on the shortest texts (question/answer pairs). Paggio (2006) reports kappa values between 0.7–0.8 for topic/focus, depending on the corpus section.

All of this shows that agreement is generally low, but some categories can be annotated more reliably than others. As mentioned above some problems are due to unclear documentation but many point to interesting theoretical areas. It is interesting to qualitatively evaluate those cases where the annotators disagree. As an example consider the final sentence in (9), taken from Cook & Bildhauer (2011). It is difficult to decide whether the fronted non- or the pronominal subject following the verb is the topic, since phrases in the German prefield and pronominal subjects are both often topics. The context does not always help: Both the *Erfahrung* 'experience' and the speaker (*ich*) are introduced in the sentence before (the latter in the prefield).

(9) *Dazugelernt habe ich besonders im Bereich der Öffentlichkeitsarbeit.*
more-learned have I especially in-the area the PR-work.
Ich merkte, welche Handlung welche Reaktion auslöst und wie man gewisse
I noticed which action which reaction causes and how one certain
Ereignisse richtig kommuniziert. Von [dieser Erfahrung]$_{TOP?}$ *kann [ich]*$_{TOP?}$
events correctly communicate From this experience can I
am neuen Ort selbstverständlich profitieren.
at new place clearly benefit
'I learned more in the area of public relations work in particular. I noticed what sort of

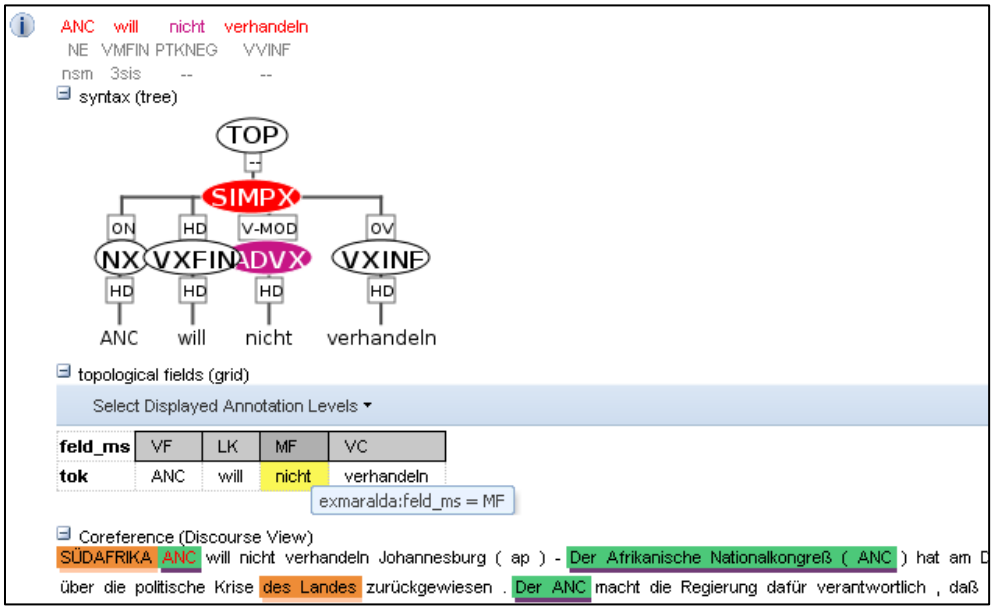
reaction was caused by which actions and how to communicate certain events correctly. I will clearly be able to benefit from this experience at the new location'

The following issues have been reported to be problematic for IS annotation: Before even assigning an IS category annotators often disagree already on the markables (units for annotation) because it is difficult to find an operationalizable definition of referential expressions (in particular in idioms, metaphors, bare plurals, spatio-temporal and numeric expressions, see Langer 2004, Cook & Bildhauer 2011). Additionally, the space between clearly given and clearly new referents is difficult to delineate. One problem is the role of lexical association or bridging (Asher & Lascarides 1998) and knowledge of prototypical situations for discourse referent accessibility (cf. Miéville 1999 and Vieira and Poesio 2000). There is no unambiguous test for deciding whether a sentence is thetic or has topic-comment structure (cf. McNally 1998), nor is there a definitive test to determine which of several candidates is the aboutness topic.

Even though annotation efforts are, at first glance, disappointing, they lead to valuable insights. To be of real value, any scientific descriptive system should either be comprehensive, i.e. applicable to any case (=utterance) or else specify the limitations of its applicability. On both these ends, work is needed on delineating problematic cases and finding appropriate solutions. The goal remains to find a satisfactory system allowing (after training) for an uncontroversial annotation *independently of the person carrying it out*, which the studies presented here all seek to establish. In this respect, annotation projects offer the best chance of bridging the gap between textbook examples and truly comprehensive theory building. While some questions remain open, studies using existing corpora show that annotatable IS categories are meaningful and useful for explaining linguistic phenomena.

3.4 Using indirect evidence

Though direct IS annotation is vulnerable to low IAA, other, more reliable, types of annotation can be used to study IS *indirectly*. These include syntactic analyses, word order, pronominalization and argument omission, properties of discourse referents and coreference, as well as rhetorical structures. In this section we would like to show how richly annotated corpora can give us insights into IS phenomena using categories that are independent from the IS annotations above. For this purpose we will use the German treebank TüBa-D/Z (Telljohann et al. 2009). Version 6, which will be used below, contains over 975,000 tokens with parts-of-speech and lemmas, constituent trees with grammatical functions and topological fields, and coreference annotation for several kinds of relations (anaphora, cataphora, apposition and more).[13] Topological fields describe the position of constituents relative to the verbs in each clause and are particularly valuable for the study of IS: the first phrase in the main clause stands in the 'prefield', before the main verb. Constituents in the prefield often contain topical information, but are also used for contrastive foci (cf. Speyer 2010).



[13] We use a version of the corpus that separates topological fields from phrase structure trees for better searchability and visualization as shown in Figure 2 (see Krause et al. 2011 for details).

Figure 2. Multilayer representation of TueBa-D/Z with separate layers for syntax, topological fields and discourse level coreference links in ANNIS, reproduced from Krause et al. (2011).

Using these annotations we can find a variety of constructions which are of interest for IS research. As an example, consider the German expletive *es* 'it', illustrated in (10).

(10) *Es tanzen Menschen auf der Strasse* 'People are dancing in the street'
it dance.PL people on the street

The construction has been described as 'thetic', marking propositions as 'all new' (see Önnerfors 1996: 305–306): the hearer presumably does not know there are people dancing in the street, nor have they been mentioned. To test whether this is true, we can combine syntactic annotation, topological fields and coreference links. We search for sentences with expletive *es* in the prefield (topological annotation and grammatical function), followed by the main verb and its subject (syntactic constituent annotation), and then check whether the subject has an antecedent of any sort (coreference annotation, which distinguishes direct anaphora, bridging and other cases not relevant here). The result is that subjects in expletive *es* sentences overwhelmingly lack an antecedent (89/91 cases, cf. Ritz et al. 2010).

At the same time, we can find unusual sentences that conflict with the generalization:

(11) [Heading:] *"Die CDU denkt da ganz anders"*
Die Ex-Kultursenatorin ***Helga Trüpel*** *(grün) und die Kulturpolitikerin* ***Carmen Emigholz*** *(rot) schauen zurück auf vier Jahre schwarz-rote Kulturpolitik* [...7 intervening sentences] ***Es*** *diskutierten die kulturpolitische Sprecherin der SPD-Fraktion* ***Carmen Emigholz*** *und die bündnisgrüne Spitzenkandidatin* ***Helga Trüpel*** *(von 1991 bis 1995 Kultursenatorin).* [Beginning of interview]
'[Heading:] "The CDU thinks completely differently"

The former culture senator **Helga Trüpel** (Green) and the culture politician **Carmen Emigholz** (Red) look back on four years of Black-Red culture politics [...7 intervening sentences] The culture political speaker of the SPD faction **Carmen Emigholz** and the Green Alliance head candidate **Helga Trüpel** (culture senator from 1991 to 1995) lead the discussion. [Beginning of interview]'

This unusual example shows just how inactive an aforementioned referent must be in order to be compatible with the expletive *es* construction. The relevant persons are mentioned at the beginning of the article just under the main heading, only to be referred to again eight sentences later (the intervening space delivers background information relevant for the interview that follows). Note also that upon reintroduction, additional information about the referents is supplied in appositions, which are typical when introducing discourse referents: we are told who these politicians are and what they have done, so that the proposition predicated on the subject, that a discussion took place, is presented as no newer than the biographical data about the politicians. In this way corpus data gives us not only insight on the strong correlation between theticity and the expletive *es* construction, but also shows us the limits of givenness: when do given/accessible referents become inaccessible? And under what circumstances may a thetic construction contain familiar referents?

A further advantage of 'naturally occurring' corpus data can be seen here: we can examine the entire context of a referent to better understand the factors involved in its syntactic realization. In fact, with the availability of coreference annotation, we need not stop at searching for antecedents, but can also investigate how thetic subjects are integrated into following discourse. In a 'file-card' metaphor of common ground management (cf. Krifka 2008: 41), we might expect thetic subjects to join the given referents elaborated on in further discourse. Searching for referents pointing back to thetic subjects in subsequent discourse, we find that subjects in only 12/91 thetic sentences discovered previously are later referred back

to. This suggests a difference between the discourse-newness of expletive *es* all-new clauses: their subjects are not only novel, but the speaker probably does not intend to talk about them further. (12)-(13) illustrate this:

(12) *In seinen Bildern selbst gibt sich der Star zugänglicher. **Es** dominieren **klare Farben**, oft zentimeterdick aufgetragen, dann wieder abgekratzt*
'In his own paintings the star presents himself more accessibly. **Clear colors** dominate, often painted centimeters-thick, then scratched off again'

(13) *Papas Aquarium stand in der Küche. "Damit bin ich aufgewachsen." **Es** kamen **Hamster, Meerschweinchen, Eidechsen, Schildkröten** dazu.*
'Dad's aquarium stood in the kitchen. "I grew up with that." **Hamsters, guinea pigs, lizards, turtles** were added to that.'

The expletive subjects are part of an elaboration on a (discourse) topic, and therefore accessible through bridging in the widest sense. However, unlike textbook examples, where bridging often serves to shift discourse from whole to parts, these deliver a one-time elaboration on an already established topic: clear colors describe the artist's technique (cf. the paraphrase 'he paints with clear colors...') and the list of pets illustrates the father's interest in animals (cf. 'dad got hamsters, guinea pigs...'). Thus while annotating IS is difficult, we can find examples of phenomena like theticity indirectly using annotations that are easier to operationalize, and discuss e.g. cases of newness with special usage patterns.

3.5 Automatic annotation of IS categories

Automatic annotation methods mean more text can be annotated with less human effort, so that hypotheses can be tested more quickly. In Computational Linguistics, automatic annotation of IS categories plays a role in computational discourse understanding. The most important step in this area is coreference resolution, since it is necessary for any discourse

understanding application processing more than single sentences, but for space reasons, we cannot give an extensive review of this vast field here (see Stede 2011, Chapter 3 for an overview). Instead we focus here on IS categories proper, though note that detecting the information status of a noun phrase can be seen as a sub-problem of coreference resolution: Given an NP, first determine its information status, and only if is not new, search for an antecedent in prior discourse.

Annotating information status automatically involves two subtasks: defining a set of features to compute for each NP and selecting a suitable machine learning (ML) algorithm. A statistical model is derived from a corpus annotated with those features and the target information (information status), which is then used assign information status labels automatically to unseen texts.[14]

Features can include phrase form (pronoun, name, etc.), determiner (definite, indefinite, possessive, none), position (sentence initial, final, other); previous mention of the head noun; and so on. Given a set of training instances, the ML algorithm determines relative weights of the features which produce optimal predictions. The model's quality depends on the power of features and their combinations to distinguish target classes. Part of the training data is held out in the training phase in order to ensure robust feature selection. The model can then be tested on the held-out data, and predicted classes are compared to the manually-assigned annotations. By inspecting misclassified instances, new features can be sought which are likely to yield improvements, whereupon the process is cyclically repeated.

To illustrate this procedure, consider (14) and a simple scheme distinguishing given and new referents. NPs are given IDs and manually-assigned target class labels:

[14] An alternative approach uses *rule-based* methods. See Nissim (2006) for a comparison of ML and rule-based approaches.

(14) [A student]$_{1;new}$ was walking [her dog]$_{2;new}$ when suddenly [the dog]$_{3;given}$ jumped into [the woods]$_{4;new}$, following [another dog]$_{5;new}$.

Assume we have automatically extracted the following for each NP:

- form (pronoun/name/other)
- determiner (definite/indefinite/possessive/none)
- position in sentence (initial/final/other)
- head mention (number of times the head noun appears in previous context)

Table 1 summarizes the input data for the ML algorithm.

ID	Form	Determiner	Position	Head-mention
1	other	indefinite	initial	0
2	other	possessive	other	0
3	other	definite	other	1
4	other	definite	other	0
5	other	none	final	2

Table 1: Sample input to the ML algorithm for (14).

Assume we have trained a model, in this case a decision tree, on large amounts of data annotated in this way. The decision tree could look like this:[15]

```
head_mention>0
    determiner=definite:given
    determiner!=definite:new
head_mention=0
```

[15] The tree comprises a series of tests: if the test in the first line yields "true", the indented lines below it are executed. Lines with a colon indicate the class to be assigned, terminating the procedure. Exclamation marks signal negation.

form=pronoun:given

form!=pronoun

determiner=none

position=initial:given

position!=initial:new

determiner=indef:new

...

This model correctly classifies NP_3 in our example as given, and all others as new.

Several attempts at automatically annotating IS categories have been made, including Nissim (2006) and Rahman & Ng (2012) for the Switchboard dialogues, Hempelmann et al. (2005) for narrative texts, Ritz (2010) and Markert et al. (2012) for OntoNotes, and Cahill & Riester (2012) for DIRNDL. While results are encouraging, generally automatic annotation of IS today is not as reliable as other annotation tasks.

Other IS notions have not received as much attention as information status. An exception is work on topic-focus articulation in Czech (Postolache et al. 2005) with a classifier using features like coreference, node position etc. For a comprehensive overview on Computational Linguistics work on IS, including coreference resolution and production of IS-marked utterances in speech synthesis, see Stede (2011).

4. Corpus-based IS studies

In this section we briefly describe principal uses corpora for IS with pointers to IS-relevant corpus studies. Most widespread is the use of corpora as ‘example banks’ to illustrate a theoretical point. While it might not be important how a naturally-occurring example is found a systematic corpus search for an IS phenomenon (using direct annotations or indirectly) is sometimes necessary, especially for low-frequency phenomena, for languages/dialects/registers that one doesn’t speak, or for contrastive studies. De Kuthy &

Meurers (2012), for example, analyze stress placement in the IMS radio news corpus to evaluate two competing models of focal intonation patterns. Going a step further, corpora can be used to systematically explore phenomena with specific searches. The most comprehensive way to explore a phenomenon is, of course, annotation and development of annotation guidelines. This is exemplified by Baumann and Riester (2012) who show that in order to understand the relationship between givenness and (de)accentuation it is necessary to distinguish between lexical and referential givenness. The study described in Cook & Bildhauer (2011) is another case in point: A systematic annotation of topic and discussion of problematic examples lead to a better understanding of different types of topics.

While exploring examples is mostly qualitative, corpora are also invaluable for quantitative studies (if the corpus design is known; see Section 32.2). Most IS studies that we are aware of use descriptive statistics. Counts are, of course, useful only in contrast – one can compare (competing) structures in the same corpus or the same structure in different corpora (Biber & Jones 2009). Breckle & Zinsmeister (2012), for example, analyze the distribution of prefield phrases with different information status in L1 and L2 German in order to find out whether learners transfer IS strategies from their native language. There are a number of papers that deal with language-change, including word-order change and development of definiteness markers triggered or influenced by IS (see the papers in Hinterhölzl & Petrova 2009, Meurmann-Solin et al. 2012 and Chapter 27). These papers are necessarily corpus-based, and some of them discuss quantitative as well as qualitative aspects of IS.

More involved inferential statistics and modeling are not (yet) common in IS studies. This will certainly change over the next few years, as more and more annotated data becomes available.

5. Summary & conclusion

Information-structure theory is concerned with the description and prediction of surface forms of utterances, based on previous discourse and participants' shared knowledge (common ground). Corpus data is often much more complex than made-up examples – it is sometimes difficult to tease apart the different types of knowledge that influence the shape of an utterance in order to see where exactly IS comes in. While it is therefore useful to work with controlled data in order to *formulate* the theoretical notions one works with, it becomes necessary at some point to look at authentic data. In this chapter we describe the use of corpora for the study of IS. We have shown what role corpus design plays and which studies can be performed with corpus data. We have dealt primarily with IS annotation, introduced multilayer architectures where different types of information can be added on independent levels and described different annotation schemes. We have discussed evaluating annotations and some of the problems that might be responsible for disagreement in annotating IS notions. The complexity of corpus data is an advantage for research rather than a problem, if corpora are well designed (for the research question) and suitably annotated, *because* corpus data does not permit 'the easy way out' – it shows us where our notions are not yet well-defined and where our models are not complex enough to explain what we want to understand.

6. References

Artstein, Ron, and Poesio, Massimo (2008). 'Inter-Coder Agreement for Computational Linguistics', *Computational Linguistics* 34(4): 556–596.

Asher, Nicholas, and Lascarides, Alex (1998). 'Bridging', *Journal of Semantics* 15(1): 83–113.

Baumann, Stefan, Brinckmann, Caren, Hansen-Schirra, Silvia, Kruijff, Geert-Jan, Kruijff-Korbayová, Ivana, Neumann, Stella, Steiner, Erich, Teich, Elke, and Uszkoreit, Hans (2004). 'The MULI Project. Annotation and Analysis of Information Structure in German and English', in *Proceedings of LREC 2004*. Lisbon, 1489–1492.

Baumann, Stefan, and Riester, Arndt (2012). 'Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects', in G. Elordieta, and P. Prieto (eds.), *Prosody and Meaning*. Berlin: Mouton de Gruyter, 119–161.

Biber, Douglas (1993). 'Representativeness in Corpus Design', *Literary and Linguistic Computing* 8(4): 243–257.

Biber, Douglas (2009). ‘Multi-Dimensional Approaches’, in A. Lüdeling, and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Vol. 2. Berlin: Mouton de Gruyter, 822–855.

Biber, Douglas, and Jones, James K. (2009). ‘Quantitative Methods in Corpus Linguistics’, in A. Lüdeling, and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Vol. 2. Berlin: Mouton de Gruyter, 1286–1304.

Bildhauer, Felix (2011). ‘Mehrfache Vorfeldbesetzung und Informationsstruktur. Eine Bestandsaufnahme’, *Deutsche Sprache* 4/11: 362–379.

Bouma, Gerlof, Øvrelid, Lilja, and Kuhn, Jonas (2010). ‘Towards a Large Parallel Corpus of Cleft Constructions’, in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta, 3585–3592.

Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang, and Smith, George (2002). ‘The TIGER Treebank’, in *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*. Sozopol, Bulgaria, 24–42.

Breckle, Margit, and Zinsmeister, Heike (2012). ‘A Corpus-Based Contrastive Analysis of Local Coherence in L1 and L2 German’, in V. Karabalić, M. A. Varga, and L. Pon (eds.), *Discourse and Dialogue*. Frankfurt am Main: Peter Lang, 235–250.

Cahill, Aoife, and Riester, Arndt (2012). ‘Automatically Acquiring Fine-Grained Information Status Distinctions in German’, in *Proceedings of the 13th Annual SIGdial Meeting on Discourse and Dialogue*. Seoul, 232–236.

Carletta, Jean (1996). ‘Assessing Agreement on Classification Tasks: The Kappa Statistic’, *Computational Linguistics* 22(2): 249–254.

Carletta, Jean, Evert, Stefan, Heid, Ulrich, Kilgour, Jonathan, Robertson, Judy, and Voormann, Holger (2003). ‘The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data’, *Behavior Research Methods, Instruments, and Computers* 35(3): 353–363.

Chiarcos, Christian, Fiedler, Ines, Grubic, Mira, Haida, Andreas, Hartmann, Katharina, Ritz, Julia, Schwarz, Anne, Zeldes, Amir, and Zimmermann, Malte (2011). ‘Information Structure in African Languages: Corpora and Tools’, *Language Resources and Evaluation* 45(3): 361–374.

Chinchor, Nancy, and Sundheim, Beth (2003). *Message Understanding Conference (MUC) 6*. Philadelphia: Linguistic Data Consortium.

Cook, Philippa, and Bildhauer, Felix (2011). ‘Annotating Information Structure. The Case of “Topic”’, in S. Dipper, and H. Zinsmeister (eds.), *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena.* (Bochumer Linguistische Arbeitsberichte 3.) Bochum: Ruhr-Universität Bochum, 45–56.

De Kuthy, Kordula, and Meurers, Detmar (2012). ‘Focus Projection between Theory and Evidence’, in S. Featherston, and B. Stolterfoth (eds.), *Empirical Approaches to Linguistic Theory - Studies in Meaning and Structure*. Berlin: De Gruyter, 207–240.

van Deemter, Kees, and Kibble, Rodger (2000). ‘On Coreferring: Coreference in MUC and Related Annotation Schemes’, *Computational Linguistics* 26(4): 629–637.

Dipper, Stefanie (2005). ‘XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation’, in *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, Germany, 39–50.

Dipper, Stefanie, Götze, Michael, and Skopeteas, Stavros (eds.) (2007). ‘Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure’, *Interdisciplinary Studies on Information Structure* 7, special issue.

Eckart, Kerstin, Riester, Arndt, and Schweitzer, Katrin (2012). ‘A Discourse Information Radio News Database for Linguistic Analysis’, in C. Chiarcos, S. Nordhoff, and S. Hellmann (eds.), *Linked Data in Linguistics*. Berlin: Springer, 65–76.

Godfrey, John J., Holliman, Edward C., and McDaniel, Jane (1992). 'SWITCHBOARD: Telephone Speech Corpus for Research and Development', in *Proceedings of ICASSP-92*. San Francisco, CA, 517–520.

Golcher, Felix (2012). *Wiederholungen in Texten. Segmentieren und Klassifizieren mit vollständigen Substringfrequenzen*. PhD Thesis, Humboldt-Universität zu Berlin.

Gries, Stefan Th. (2006). 'Exploring Variability within and between Corpora: Some Methodological Considerations', *Corpora* 1(2): 109–151.

Hajič, Jan, Panevová, Jarmila, Hajičová, Eva, Panevová, Jarmila, Sgall, Petr, Pajas, Petr, Štěpánek, Jan, Havelka, Jiří, and Mikulová, Marie (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

Hajičová, Eva, Panevová, Jarmila, and Sgall, Petr (2000). 'Coreference in Annotating a Large Corpus', in *Proceedings of LREC-2000*. Athens, 497–500.

Haug, Dag T.T., Eckhoff, Hanne M., Majer, Marek, and Welo, Eirik (2009). 'Breaking Down and Putting Back Together: Analysis and Synthesis of New Testament Greek', *Journal of Greek Linguistics* 9(1): 56–92.

Haug, Dag T. T., Eckhoff, Hanne Martine, and Welo, Eirik (to appear). 'The Theoretical Foundations of Givenness Annotation', in K. Bech, and K. Eide (eds.), *Information Structure and Syntax in Germanic and Romance Languages*. Amsterdam: John Benjamins.

Hempelmann, Christian F., Dufty, David, McCarthy, Philip M., Graesser, Arthur C., Cai, Zhiqiang, and McNamara, Danielle S. (2005). 'Using LSA to Automatically Identify Givenness and Newness of Noun-Phrases in Written Discourse', in B. Bara (ed.), *Proceedings of the 27th Annual Meetings of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, 941–946.

Hendrickx, Iris, Boumaz, Gosse, Coppens, Frederik, Daelemans, Walter, Hoste, Veronique, Kloostermanz, Geert, Mineurz, Anne-Marie, Vloet, Joeri Van Der, and Verschelde, Jean-Luc (2008). 'A Coreference Corpus and Resolution System for Dutch', in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakech, 144–149.

Hinterhölzl, Roland, and Petrova, Svetlana (eds.) (2009). *Information Structure and Language Change. New Approaches to Word Order Variation in German*. Berlin: Mouton de Gruyter.

Hirschman, Lynette, Robinson, Patricia, Burger, John D., and Vilain, Marc B. (1998). *Automating Coreference: The Role of Annotated Training Data*. AAAI technical report SS-98-01. Available at: http://www.aaai.org/Papers/Symposia/Spring/1998/SS-98-01/SS98-01-018.pdf.

Hovy, Eduard, Marcus, Mitchell, Palmer, Martha, Ramshaw, Lance, and Weischedel, Ralph (2006). 'OntoNotes: The 90% Solution', in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York: Association for Computational Linguistics, 57–60. Available at: http://www.aclweb.org/anthology/N/N06/N06-2015.

Johansson, Mats (2001). 'Clefts in Contrast: A Contrastive Study of it Clefts and wh Clefts in English and Swedish Texts and Translations', *Linguistics* 39(3): 547–582.

Kilgarriff, Adam (2012). 'Getting to Know Your Corpus', in P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.), *Text, Speech and Dialogue. 15th International Conference, TSD 2012, Brno, Czech Republic*. (Lecture Notes in Computer Science 7499.) Berlin and Heidelberg: Springer, 3–15.

Komagata, Nobo (1999). *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese*. PhD Thesis, University of Pennsylvania.

Krause, Thomas, Ritz, Julia, Zeldes, Amir, and Zipser, Florian (2011). 'Topological Fields, Constituents and Coreference: A New Multi-layer Architecture for TüBa-D/Z', in H. Hedeland, T. Schmidt, and K. Wörner (eds.), *Multilingual Resources and Multilingual*

Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011. (Working Papers in Multilingualism 96.) Hamburg: Universität Hamburg, 259–262.
Krifka, Manfred (2008). 'Basic Notions of Information Structure', *Acta Linguistica Hungarica* 55: 243–276.
Krippendorff, Klaus (1980). *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.
Labov, William (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
Langer, Stefan (2004). 'A Linguistic Test Battery for Support Verb Constructions', *Verbes Supports. Nouvel état des lieux. Special issue of Linguisticae Investigationes* 27(2): 171–184.
Lüdeling, Anke (2011). 'Corpora in Linguistics: Sampling and Annotation', in K. Grandin (ed.), *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. (Nobel Symposium 147.) New York: Science History Publications, 220–243.
Lüdeling, Anke, and Kytö, Merja (eds.) (2008-2009). *Corpus Linguistics. An International Handbook*. (Handbooks of Linguistics and Communication Science 29.) Berlin and New York: Mouton de Gruyter.
Mann, William C., and Thompson, Sandra A. (1988). 'Rhetorical Structure Theory: Toward a Functional Theory of Text Organization', *Text* 8(3): 243–281.
Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann (1993). 'Building a Large Annotated Corpus of English: The Penn Treebank', *Special Issue on Using Large Corpora, Computational Linguistics* 19(2): 313–330.
Markert, Katja, Hou, Yufang, and Strube, Michael (2012). 'Collective Classification for Fine-Grained Information Status', in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea, 795–804.
McNally, Louise (1998). 'On Recent Formal Analyses of Topic', in J. Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Lévy, and E. Vallduví (eds.), *The Tbilisi Symposium on Language, Logic and Computation: Selected Papers*. Stanford: CSLI Publications.
Meurers, Detmar, Ott, Niels, and Ziai, Ramon (2010). 'Compiling a Task-Based Corpus for the Analysis of Learner Language in Context', in *Pre-Proceedings of Linguistic Evidence 2010*. Tübingen, 214–217.
Meurmann-Solin, Anneli, López-Couso, María José, and Los, Bettelou (eds.) (2012). *Information Structure and Syntactic Change in the History of English*. Oxford: Oxford University Press.
Miéville, Denis (1999). 'Associative Anaphora: An Attempt at a Formalization', *Journal of Pragmatics* 31: 327–337.
Müller, Christoph, and Strube, Michael (2006). 'Multi-Level Annotation of Linguistic Data with MMAX2', in S. Braun, K. Kohn, and J. Mukherjee (eds.), *Corpus Technology and Language Pedagogy*. Frankfurt: Peter Lang, 197–214.
Naumann, Karin (2006). *Manual for the Annotation of In-Document Referential Relations*. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen. Available at: http://arbuckle.sfs.uni-tuebingen.de/resources/tuebadz_relations_man.pdf.
Nissim, Malvina (2006). 'Learning Information Status of Discourse Entities', in *Proceedings of the 2006 Conference on Emprical Methods in Natural Language Processing (EMNLP 2006)*. Sydney, Australia, 94–102.
Nissim, Malvina, Dingare, Shipra, Carletta, Jean, and Steedman, Mark (2004). 'An Annotation Scheme for Information Status in Dialogue', in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, Portugal, 1023–1026.

Önnerfors, Olaf (1996). 'On Narrative Declarative V1 Sentences in German', in T. Swan, and O. J. Westvik (eds.), *Modality in Germanic Languages: Historical and Comparative Perspectives*. (Trends in Linguistics: Studies and Monographs 99.) Berlin: Mouton de Gruyter, 293–319.

Ott, Niels, Ziai, Ramon, and Meurers, Detmar (2012). 'Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context', in T. Schmidt, and K. Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*. (Hamburg Studies in Multilingualism 14.) Amsterdam: John Benjamins, 47–69.

Paggio, Patrizia (2006). 'Annotating Information Structure in a Corpus of Spoken Danish', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. Genova, Italy, 1606–1609.

Petrova, Svetlana, Solf, Michael, Ritz, Julia, Chiarcos, Christian, and Zeldes, Amir (2009). 'Building and Using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian', *Traitement automatique des langues* 50(2): 47–71.

Poesio, Massimo (2004a). 'Discourse Annotation and Semantic Annotation in the GNOME Corpus', in *Proceedings of the ACL Workshop on Discourse Annotation*. Barcelona, 72–79.

Poesio, Massimo (2004b). 'The MATE/GNOME Scheme for Anaphoric Annotation, Revisited', in M. Strube, and C. Sidner (eds.), *Proceedings of SIGDIAL*. Boston, 154–162.

Poesio, Massimo, and Artstein, Ron (2005). 'The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account', in *Proceedings of ACL Workshop on Frontiers in Corpus Annotation*. Stroudsburg, PA: Association for Computational Linguistics, 76–83.

Poesio, Massimo, and Artstein, Ron (2008). 'Anaphoric Annotation in the ARRAU Corpus', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakech, 1170–1174.

Poesio, Massimo, Bruneseaux, Florence, and Romary, Laurent (1999). 'The MATE Meta-Scheme for Coreference in Dialogues in Multiple Languages', in *Proceedings of the ACL Workshop on Standards for Discourse Tagging*. College Park, MD, 65–74.

Poesio, Massimo, and Vieira, Renata (1998). 'A Corpus-Based Investigation of Definite Description Use', *Computational Linguistics* 24(2): 183–216.

Postolache, Oana, Kruijff-Korbayová, Ivana, and Kruijff, Geert-Jan (2005). 'Data-Driven Approaches for Information Structure Identification', in *Proceedings of the Conference on Emperical Methods in Natural Language Processing*. Vancouver, 9–16.

Powers, David M. W. (2012). 'The Problem with Kappa', in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*. Avignon, France, 345–355.

Prince, Ellen F. (1981). 'Toward a Taxonomy of Given-New Information', in P. Cole (ed.), *Radical Pragmatics*. New York: Academic Press, 223–255.

Rahman, Altaf, and Ng, Vincent (2012). 'Learning the Fine-Grained Information Status of Discourse Entities', in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*. Avignon, France, 798–807.

Riester, Arndt, Killmann, Lorena, Lorenz, David, and Portz, Melanie (2007). *Richtlinien zur Annotation von Gegebenheit und Kontrast in Projekt A1. Draft version, November 2007*. Technical Report, SFB 732, University of Stuttgart, Stuttgart, Germany.

Riester, Arndt, Lorenz, David, and Seemann, Nina (2010). 'A Recursive Annotation Scheme for Referential Information Status', in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta, 717–722.

Ritz, Julia (2010). 'Using tf-idf-related Measures for Determining the Anaphoricity of Noun Phrases', in *Proceedings of KONVENS 2010*. Saarbrücken, 85–92.

Ritz, Julia, Chiarcos, Christian, and Bieler, Heike (2010). 'On the Information Structure of Expletive Sentences: An Empirical Study across Multiple Layers of Annotation', in *32. Jahrestagung der deutschen Gesellschaft für Sprachwissenschaft, Sektion CL*. Berlin, 315.

Ritz, Julia, Dipper, Stefanie, and Götze, Michael (2008). 'Annotation of Information Structure: An Evaluation Across Different Types of Texts', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakech, 2137–2142.

Ritz, Julia (to appear). *Discourse-Givenness of Noun Phrases - Theoretical and Computational Models*. PhD Thesis, Universität Potsdam.

Rooth, Mats (1992). 'A Theory of Focus Interpretation', *Natural Language Semantics* 1: 75–116.

Schmidt, Thomas, and Wörner, Kai (2009). 'EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research', *Pragmatics* 19(4): 565–582.

Skopeteas, Stavros, Fiedler, Ines, Hellmuth, Sam, Schwarz, Anne, Stoel, Ruben, Fanselow, Gisbert, Féry, Caroline, and Krifka, Manfred (2006). 'Questionnaire on Information Structure (QUIS)', *Interdisciplinary Studies on Information Structure* 4, special issue.

Speyer, Augustin (2010). 'Filling the German Vorfeld in Written and Spoken Discourse', in S.-K. Tanskanen, M.-L. Helasvuo, M. Johansson, and M. Raitaniemi (eds.), *Discourses in Interaction*. (Pragmatics & Beyond New Series 203.) Amsterdam and Philadelphia: John Benjamins, 263–290.

Stede, Manfred (2004). 'The Potsdam Commentary Corpus', in B. Webber, and D. K. Byron (eds.), *Proceeding of the ACL-04 Workshop on Discourse Annotation*. Barcelona, Spain, 96–102.

Stede, Manfred (2011). *Discourse Processing*. (Synthesis Lectures on Human Language Technologies 4.) [San Rafael, CA]: Morgan & Claypool.

Telljohann, Heike, Hinrichs, Erhard W., and Kübler, Sandra (2003). *Stylebook for the Tübingen Treebank of Written German*. Universität Tübingen, Seminar für Sprachwissenschaft.

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

Versley, Yannick (2006). 'Disagreement Dissected: Vagueness as a Source of Ambiguity in Nominal (Co-)Reference', in *Proceedingsof Ambiguity in Anaphora ESSLLI Workshop*. Málaga, 83–89. Available at: http://www.versley.de/esslli06.pdf.

Vieira, Renata, and Poesio, Massimo (2000). 'Corpus-based Development and Evaluation of a System for Processing Definite Descriptions', in *Proceedings of the 18th conference on Computational linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 899–903. Available at: http://dx.doi.org/10.3115/992730.992776.

Weischedel, Ralph, Pradhan, Sameer, Ramshaw, Lance, Micciulla, Linnea, Palmer, Martha, Xue, Nianwen, Marcus, Mitchell, Taylor, Ann, Babko-Malaya, Olga, Hovy, Eduard, Belvin, Robert, and Houston, Ann (2007). *OntoNotes Release 1.0*. Philadelphia: Linguistic Data Consortium.

Weskott, Thomas, Hörnig, Robin, Fanselow, Gisbert, and Kliegl, Reinhold (2011). 'Contextual Licensing of Marked OVS Word Order in German', *Linguistische Berichte* 225: 3–18.

Zeldes, Amir, Ritz, Julia, Lüdeling, Anke, and Chiarcos, Christian (2009). 'ANNIS: A Search Tool for Multi-Layer Annotated Corpora', in *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.