

# **Beyond semantics: the challenges of annotating pragmatic and discourse phenomena**

## **Introduction to the special issue**

**Stefanie Dipper**

*Institute of Linguistics  
Ruhr-University Bochum*

DIPPER@LINGUISTICS.RUB.DE

**Heike Zinsmeister**

*Institute of German Language and Literature  
University of Hamburg*

HEIKE.ZINSMEISTER@UNI-HAMBURG.DE

**Bonnie Webber**

*Institute for Language, Cognition and Computation  
University of Edinburgh*

BONNIE@INF.ED.AC.UK

### **1. Introduction**

Manual corpus annotation and automated corpus analysis began with a focus on what was seen as “low-hanging fruit”, such as morphological tags, part-of-speech tags, and syntactic structures. After investment of considerable resources (time, money and effort), development of fairly accurate automated tools, and recognition that other slightly higher-hanging fruit were there for the taking, attention shifted—this time, to characterizing the meaning and intention of language use, including the development of semantic labelling schemes such as word sense tagging (e.g. Miller et al., 1993), semantic class labelling such as named-entity typing (MUC-6, 1995; Brunstein, 2002; ACE-EDT, 2004), semantic role labelling (Palmer et al., 2005; Fillmore et al., 2004; Meyers et al., 2004), coreference linking (Passonneau, 1996; Hirschman and Chinchor, 1998; Poesio, 2000) and temporal relations (Pustejovsky et al., 2003a,b) and the development of pragmatic labelling schemes such as dialogue act tagging (Allen and Core, 1997; Carletta et al., 1997; Jurafsky et al., 1997; Alexandersson et al., 1998). These too have seen considerable resources invested in them, as well as the development of fairly accurate automated tools.

But horizons change, and attention is now focussed somewhat higher—on problems in understanding and reliably annotating other sorts of semantic and pragmatic phenomena, especially phenomena that involve spans of text larger than a single sentence or clause.

The goal of this special issue is to show the challenges faced in reliably annotating abstract semantic and pragmatic information at both the sentence and discourse levels, and how those challenges are being met. Such information is frequently not explicitly or unambiguously marked in natural language. It is usually dependent on contextual information, and annotators often have to reconstruct complex relations and situations from the context.

Annotated data can serve both as the basis of linguistic investigations and as training data for applications developed in the field of natural language processing. Most of the papers in this issue

deal with resources that are intended to ultimately serve automatic applications. Still, they have a strong focus on the linguistic foundations underlying the annotations.

## 2. The Challenges

Linguistically motivated semantic and pragmatic theories are too often based on toy examples, well-controlled contexts, and the intuitions of theory’s author. As a consequence, it is often hard to transfer results from theoretical linguistics to naturally-occurring texts.

Annotation, either by trained experts or untrained workers, is felt to be a way of dealing with semantic and pragmatic phenomena in naturally-occurring text. But as noted, annotating such phenomena often requires annotators to infer from context, complex relations and situations. Pertinent examples from the papers in this issue are:

- Constructing (low- or high-level) *questions under discussion* to determine the focus of a sentence or the discourse topic (see the articles by Riester & Baumann, and by Versley & Gastel)
- Judging whether a sentence makes broad statements about a topic or provides details (see the article by Louis & Nenkova)
- Judging whether some sentence serves to introduce a new entity or to present a situation as a whole (see the article by Cook & Bildhauer)
- Disambiguating discourse relations that are either triggered by connectives which are lexically ambiguous, or that are not lexically expressed at all (see the articles by Versley & Gastel, by Cartoni, Zufferey & Meyer, and by Hardt)
- Deciding whether two sentences are coherently related when there is no overt connective or other surface clue, and coherence is achieved only by inference based on word knowledge (see the article by Burstein, Tetreault & Chodorow)
- Deciding what speakers “do” with language as they interact across varied communication situations (see the articles by Tenbrink, Eberhard, Shi, Kübler & Scheutz, and by Morgan, Oxley, Bender, Zhu, Gracheva & Zachry)
- Discriminating between different lexically triggered inference relations such as presupposition and logical entailment (see the article by Tremper & Frank)

Even with explicit annotation guidelines, discourse-related and pragmatic phenomena are often difficult to annotate reliably. For instance, earlier studies (Ritz et al., 2008) showed that annotating information-structural features in German often results in inter-annotator agreement scores well below  $\kappa = 0.6$  (Cohen, 1960)—such scores are often assumed to allow for tentative conclusions only (Landis and Koch, 1977). Similarly, the overview on inter-annotator agreement measures by Artstein and Poesio (2008) showed that annotation of discourse-related features, such as dialogue act tagging, discourse segmentation, or word sense tagging, also achieves low  $\kappa$  scores in many studies.

A possible approach to tackling these problems is the use of proxies for more abstract linguistic concepts in terms of surface clues that are more reliably classified by annotators than the original concepts. A prominent example is the practice applied in the Penn Discourse Treebank (Prasad

et al., 2008), where annotators are asked to generate overt connectives for otherwise unmarked discourse relations. Another approach is to use paraphrase tests to elicit interpretations in a way as objective as possible (cf. Zhou and Xue, 2012). Pertinent examples from the current papers are:

- Inserting overt connectives to determine discourse relations (see the article by Versley & Gastel)
- Approximating meaning of ambiguous connectives by means of their translation equivalents (see the article by Cartoni, Zufferey & Meyer)
- Transforming a sentence into *Concerning X, ...* to test whether X is an aboutness topic (see the article by Cook & Bildhauer)
- Paying attention to morphological endings as indicators of nominal clauses that have predicative potential and can function as arguments of discourse relations (see the article by Zeyrek, Demirşahin, Sevdik Callı & Cakıcı)
- Guiding the annotators by a decision-tree like question-based annotation design to elicit complex semantico-pragmatic judgements in a reliable way (see the article by Tremper & Frank)
- Taking the reverse route by explicitly annotating the entities that appear to signal coherence in a corpus that already contains coherence annotation (see the article by Taboada & Das).

### 3. The Case for a Special Issue

The general idea of the special issue was to gather research that reports on the generation (and exploitation) of corpora that are annotated with pragmatic or discourse-related information grounded in linguistic theory.

The volume aims at enhancing mutual awareness of people working on different kinds of abstract semantic and pragmatic phenomena from different perspectives. We would like to bring 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 special issue is to enhance exchange and awareness between researchers of both fields, and to gain insights in the—possibly common—properties and peculiarities of these abstract semantic and pragmatic phenomena. Ideally, the volume would allow people to realize that there are problems they share—despite the fact that they are working on quite different tasks—and to recognize (partial) solutions that they too might be able to adopt.

We also see it as an important desideratum to promote the application of linguistic theories to naturally-occurring texts. This would enhance the search for operationalizations of theoretical concepts, which can probably then be annotated with higher reliability. It would open up corpus-based development and validation of theoretical hypotheses. At the same time, operationalized theoretical concepts and reliable annotations would facilitate the use of pragmatic and discourse-related knowledge in computational linguistics.

This means, on one hand, that we need more theoretical linguists annotating corpora and validating their theories based on corpora, and, on the other hand, more computational linguists drawing from linguistic insights to a greater extent when annotating training data.

We hope that this special issue promotes this exchange considerably, by highlighting relevant research.

## 4. Overview of the papers

The papers collected in this special volume address topics from different fields: discourse relations and coherence, dialogue acts, text specificity and communicative goals, inference-triggering relations, and information structure.

### 4.1 Discourse relations and coherence

The paper by **Yannick Versley & Anna Gastel: “Linguistic Tests for Discourse Relations in the TüBa-D/Z Corpus of Written German”**, deals with the annotation of discourse relations in a German corpus. Annotators first segment the texts in topic segments, which answer a high-level “question under discussion”; annotators are also asked to make the topic explicit. Discourse relations usually occur within the boundaries of a topic segment. Exceptions are cross-topic relations, which are introduced by the authors to attenuate the effects of topic shifts. Relations can be subordinating or coordinating, and fall into five groups: contingency, expansion, temporal, comparison, reporting. They are assigned by means of linguistic tests, such as substitution or insertion of connectives, use of paraphrases or nominalizations, or explicit insertion of the questions under discussion.

The article by **Bruno Cartoni, Sandrine Zufferey & Thomas Meyer: “Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique”** focuses on disambiguating the senses of discourse connectives without annotating their arguments. The paper introduces a method of sense disambiguation based on projecting meaning cross-linguistically in a parallel corpus. They point out that connectives like *while* are ambiguous and introduce different discourse relations depending on the context (and sometimes even simultaneously). *While*, for example, has among others a *concessive* and a *contrastive* reading, in addition to its *temporal* reading. The authors highlight that the distinctions of some relations result in very low inter-annotator agreement. The method they suggest instead of sense disambiguation requires the annotators to identify the translation equivalent of the connective in the aligned sentence (particle, paraphrase, or no translation). To cluster monolingual equivalent classes of connectives, the authors conducted a cloze/fill-the-blank test in which annotators had to choose a connective from a list of connectives to fill the blank in a sentence. The authors trained a Maximum Entropy classifier to determine the senses of the connective *while* automatically (six senses). Finally the authors discuss that the cross-linguistic projection method helps to identify sub-senses in connectives not explicitly distinguished in the source language.

**Deniz Zeyrek, Işın Demirşahin, Ayışığı Sevdik Callı & Ruket Cakıcı** report in their paper **“Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language”** on the annotation of explicit discourse connectives in the Turkish Discourse Bank. They detail the annotation cycle and provide an overview of the finalized annotation scheme that has been strongly influenced by the annotation scheme of the English Penn Discourse Treebank (PDTB). Extensions and modifications implemented in their scheme concern the tag for shared discourse arguments, and the *modifier* and the *supplementary* tags of the PDTB scheme. Nominalizations are also annotated as discourse arguments, given their frequency of occurrence in Turkish. The evaluation of inter-annotation agreement shows that the task is more feasible for certain discourse relations and more subjective for others. The authors also provide statistical information about the connectives that proved challenging in the annotation process, namely discourse adverbials, subordinators and their polymorphous occurrences.

The paper by **Daniel Hardt: “A Uniform Syntax and Discourse Structure: the Copenhagen Dependency Treebanks”**, introduces the annotation of the Danish Copenhagen Dependency Treebank (CDT), which belongs to a word-aligned multi-lingual and multi-level annotated treebank comprising Danish texts and their translations into English, German, Italian, and Spanish subcorpora. Even if the CDT discourse annotation is heavily inspired by the English Penn Discourse Treebank (PDTB) annotation guidelines, the annotation procedure of the CDT differs from the PDTB by annotating on top of syntactic (dependency) structure rather than raw text strings. In this paper, the author argues from a semantic perspective for doing so, by discussing cases of contrastive, causal, and conjunctive relations.

**Maite Taboada & Debopam Das** in their article “**Annotation upon Annotation: Adding Signalling Information to a Corpus**” take the reverse way: They start from an already existing discourse treebank, the RST Discourse Treebank (Carlson et al., 2003), and mark linguistic expressions that could serve as indicators of the annotated relations—a case of reverse engineering. In their pilot study, they annotate various kinds of linguistic expressions that can signal discourse relations. The signals include discourse markers, relations between entities (e.g. coreference), semantic relations (e.g. antonyms, hypernyms), etc. Their analysis finds 14% of the relations in the RST Discourse Treebank to lack an overt signal and only 22% of the overtly signalled relations to be signalled by a discourse marker. Many relations that are signalled by other means are redundantly marked, i.e. by several signals.

**Jill Burstein, Joel Tetreault & Martin Chodorow** deal in their contribution “**Holistic Annotation of Discourse Coherence Quality in Noisy Essay Writing**” with the annotation of discourse coherence in terms of scoring student essays on text-level with a two-level scale (*low coherence*, *high coherence*). They started out with a three-level annotation scale that distinguished high coherence proper from *essentially coherent* in the sense of that “text meaning can basically be constructed, but one or two identifiable points were confusing”—the annotators were also asked to mark these points of coherence breakdown. Since inter-annotator agreement for the middle score was low, the authors decided to combine it with the high coherence score in the end. These two-level scores for discourse coherence correlated well with task-independent expert essay ratings on a 5- or 6-level scale. Finally, the authors used the annotated data to train a decision-tree classifier to distinguish low and high coherence essays in a corpus of 1,555 essays. The best performing system out-performed base-line systems in all but one of ten subcorpora.

## 4.2 Dialogue acts

The paper by **Jonathan Morgan, Meghan Oxley, Emily Bender, Liyi Zhu, Varya Gracheva & Mark Zachry: “Are We There Yet?: The Development of a Corpus Annotated for Social Acts in Multilingual Online Discourse”** is concerned with multi-party discourse. The authors created two multi-lingual corpora (English, Mandarin, Russian) of computer-mediated communication in which they annotated dialogue acts in terms of “social acts”. In particular, they annotated authority claims such as contributors giving reference to their “education, training, or a history of work in an area” and positive/negative interpersonal alignment moves such as explicit agreement when taking the turn by stating “*Exactly*.”. In cases like this, it turned out to be difficult to identify sarcasm reliably. After outlining the corpus sampling from editors’ discussions of Wikipedia articles and from (written) chat discussions, the authors describe their iterative annotation process in detail. Among others, they had been monitoring annotation quality by performing a longitudinal inter-annotator

agreement study. Finally, the authors discuss some analyses based on the annotated corpora, which identify interactions among social acts, and between participant status and social acts in the data.

The article by **Thora Tenbrink, Kathleen Eberhard, Hui Shi, Sandra Kübler & Matthias Scheutz: “Annotation of negotiation processes in joint action dialogues”** discuss another type of dialogue act annotation in terms of annotating activity coordination and goal negotiation processes, as well as belief states of the participants in dialogues that are conducted in course of a joint action. The prototypical example of such a joint action dialogue is the Edinburgh Map Task corpus (Anderson et al., 1991). One important characteristic of such dialogues is that they are grounded in terms of visual and other non-linguistic cues that also influence the dialogue structure. Tenbrink et al. present a kind of reference paper for the annotation of this type of multi-modal dialogue. They give a comprehensive overview of existing schemes and introduce relevant corpora and their annotation schemes, including their own corpora. Finally, the authors point to four layers of annotation which are not always captured in existing schemes but which they argue are essential to capture the characteristic properties of such dialogues: intonation, gestures, perception of the talk domain, and task-relevant actions.

#### 4.3 Text specificity and communicative goals

The article by **Annie Louis & Ani Nenkova: “A corpus of science journalism for analyzing writing quality”**, introduces a new type of corpus consisting of science articles of different writing qualities from the New York Times. “Great” articles are those that have been chosen for the “Best American Science Writing” annual anthologies. “Very good” articles are further articles written by the top authors, and “typical” articles are articles from other authors. The corpus is divided in subcorpora of topically-related articles. The authors hypothesize that text generality/specificity and communicative goals can help distinguishing between top and average writing. In a crowd annotation, five turkers per sentence were asked to mark isolated sentences as “general” or “specific”. 64% of the sentences were assigned the same class by at least four judges. Specific sentences predominate, and tend to occur in blocks. The authors then trained a classifier for specificity, based on lexical and non-lexical features, which outperformed the baseline considerably. It turns out that confidence from the classifier is correlated with agreement among the annotators. Next, a classifier for distinguishing “great” and “very good” articles was trained. Finally, the authors investigate the communicative goals of individual sentences, and exploit syntactic similarity for automatically identifying such goals, and for judging writing quality from them.

#### 4.4 Inference-triggering relations

The paper **Galina Tremper & Anette Frank: “A Discriminative Analysis of Fine-Grained Semantic Relations including Presupposition: Annotation and Classification”** presents a corpus-based induction study of semantic relations between verbs. The relations under investigation go beyond lexical semantic relations like synonymy and hyperonymy established, for example, in WordNet. The focus of this paper is on relations between verbs that trigger inferences, in particular *presupposition*, *logical entailment*, *temporal inclusion*, *antonymy*, and *synonymy*. The authors outline a discriminative analysis of these semantic relations, which draws on the negation test that is well established in semantic and pragmatic literature. In this paper, the authors concentrate on type-level discrimination of the relations, which will be the basis for automatically deriving implicit meaning from text in future work. The authors discuss the guidelines for manual annotation and present the

results of manually annotating their gold standard of semantic relations. In addition, they also report results for their automatic classification experiments, which outperform their baseline by a large margin.

#### 4.5 Information structure

The paper by **Philippa Cook & Felix Bildhauer: “Identifying ‘aboutness topics’: two annotation experiments”** presents two experiments on annotating aboutness topics in naturally-occurring data from German. The first experiment uses Götze et al.’s (2007) guidelines, and involves annotating sentences with the main verbs *geraten*, *reagieren*, *profitieren*, *herrschen* ‘get caught, react, profit, reign’. Motivated by the (rather poor) results of inter-annotator agreement, a second annotation experiment is carried out using refined guidelines. In contrast to Götze et al.’s guidelines, they distinguish between entity-central (presentational) and event-central (event-reporting) thematic (i.e. all-thematic) expressions. The new guidelines result in better (but still low) agreement. The authors hypothesize that high agreement on a sentence indicates a prototypical case, instantiating all typical properties of aboutness-topics. To achieve higher inter-annotator agreement, the authors argue for language-specific rough-and-ready distinctions in the guidelines.

The paper by **Arndt Riester & Stefan Baumann: “Focus Triggers and Focus Types from a Corpus Perspective”** presents an integrated analysis of novelty (new information) focus and contrastive focus: Both serve to answer (implicit or explicit) questions. With contrastive focus, the hearer is able to name (at least) one contrastive alternative; with novelty focus, the alternatives remain anonymous. The authors present the RefLex annotation scheme for information status, marking the given–new distinction at lexical and referential levels. For instance, in a sequence like *A man came in. The man coughed.*, the second occurrence of the word *man* is lexically given because it is a repetition, and the phrase *the man* is referentially given because it is coreferent with the phrase *a man*. Novelty focus (information focus) can be identified based on lexical and referential givenness. Identifying contrastive focus requires information about contrastive alternatives. The authors propose an annotation scheme for alternative-eliciting features, marking cases involving focus-sensitive particles, overtly contrastive expressions, comparative constructions, etc. Finally, the authors address the issue of secondary foci in general, and second occurrence focus in particular.

#### Acknowledgments

This special issue grew out of the workshop “Beyond Semantics” at the annual conference of the German Linguistic Society in February 2011, organized by two of the editors (Dipper and Zinsmeister, 2011). We received 32 abstracts in response to our open call for statements of intent; seven of them were based on papers presented at the workshop. Among these, we invited sixteen to submit full regular papers, and seven to submit short notes. We finally received sixteen submissions.

After the reviewing process, three papers were rejected, one withdrawn, which left us with twelve papers (ten regular papers and two short notes). In general, papers were reviewed by three reviewers.

We are very grateful to the following reviewers: John Bateman, Kristy Boyer, Özlem Çetinoğlu, Jennifer Chu-Carroll, Micha Elsner, David Elson, Cathrine Fabricius-Hansen, Raquel Fernández, Dilek Hakkani-Tür, Klaus von Heusinger, Nancy Hedberg, Graeme Hirst, Hans Kamp, Ruth Kempson, Ralf Klabunde, Anke Holler, Shinichiro Ishihara, Alistair Knott, Valia Kordoni, Alan Lee, Katja Markert, Roland Meyer, Marie-Francine Moens, Malvina Nissim, Rainer Osswald, Lilja Øvrelid,

Alexis Palmer, Paul Portner, Christopher Potts, Gisela Redeker, Craig Roberts, Josef Ruppenhofer, Ted Sanders, Kiril Simov, Wilbert Spooren, Caroline Sporleder, Manfred Stede, Amanda Stent, Angelika Storrer, Elke Teich, Sara Tonelli, Carla Umbach, Yannick Versley, Thomas Weskott, Janyce Wiebe, Nianwen Xue.

Finally, we would like to thank the managing editors of *Dialogue & Discourse*, in particular, Jonathan Ginzburg (Editor-in-Chief) and Raquel Fernández for their support throughout the preparation of this issue.

## References

- ACE-EDT. Annotation guidelines for entity detection and tracking (EDT). Linguistic Data Consortium, 2004. URL <http://catalog.ldc.upenn.edu/docs/LDC2005T09/guidelines/EnglishEDTV4-2-6.PDF>. Version 4.2.6 200400401.
- Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue acts in VERBMOBIL-2 (Second edition). Verbmobil Report 226, Saarland University, Saarbrücken, Germany, 1998. URL <http://www.coli.uni-sb.de/publikationen/softcopies/Alexandersson:1998:DAV.pdf>.
- James Allen and Mark Core. Draft of DAMSL: Dialogue act markup in several layers. Technical report, University of Rochester, 1997. URL <http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/>.
- Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jaqueline C. Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4): 351–366, 1991.
- Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics (survey article). *Computational Linguistics*, 34(4):555–596, 2008.
- Ada Brunstein. Annotation guidelines for answer types. Technical report, BBN Technologies, 2002. URL <http://catalog.ldc.upenn.edu/docs/LDC2005T33/BBN-Types-Subtypes.html>.
- Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, 1997.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.



- Stefanie Dipper and Heike Zinsmeister, editors. *Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena. Proceedings of the DGfS Workshop, Göttingen*, volume 3 of *BLA (Bochumer Linguistische Arbeiten)*. Institute of Linguistics; Ruhr-University Bochum, 2011. URL <http://www.linguistics.ruhr-uni-bochum.de/bla/>.
- Charles Fillmore, Josef Ruppenhofer, and Collin F. Baker. FrameNet and representing the link between semantic and syntactic relations. In Chu-Ren Huang and Winfried Lenders, editors, *Computational Linguistics and Beyond*, Language and Linguistics Monographs Series B. Frontiers in Linguistics I, pages 19–62. Institute of Linguistics, Academia Sinica, Taipei, 2004.
- 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, *Information Structure in Cross-Linguistic Corpora*, number 7 in Interdisciplinary Studies on Information Structure (ISIS), pages 147–187. Universitätsverlag Potsdam, 2007.
- Lynette Hirschman and Nancy Chinchor. MUC-7 coreference task definition—version 3.0. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, 1998.
- Dan Jurafsky, Liz Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation. Coders manual, draft 13. Technical Report RT 97-02, University of Colorado at Boulder & SRI International, 1997. URL <http://www.stanford.edu/~jurafsky/ws97/manual.august1.html>.
- J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank project: An interim report. In *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, 2004.
- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology (HLT-93)*, pages 303–308, Stroudsburg, Pennsylvania, 1993.
- MUC-6. Named Entity task definition. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 317–322, Columbia, Maryland, 1995. URL <http://aclweb.org/anthology/M/M95/M95-1024.pdf>. Version 2.1.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, 2005.
- Rebecca J. Passonneau. Instructions for Applying Discourse Reference Annotation for Multiple Applications (DRAMA). Unpublished. Department of Computer Science, Columbia University, 1996.
- Massimo Poesio. MATE dialogue annotation guidelines: Coreference. MATE Deliverable D2.1, pages 134–187, 2000. URL <http://www.andreasmengel.de/pubs/mdag.pdf>.

- 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'08)*, pages 2961–2968, Marrakech, Morocco, 2008.
- James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. TimeML: Robust specification of event and temporal expressions in text. In Mark T. Maybury, editor, *New directions in Question Answering. Papers from the 2003 AAAI Spring Symposium*, pages 28–34. AAAI Press, Menlo Park, California, 2003a.
- James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*, pages 647–656, 2003b.
- Julia Ritz, Stefanie Dipper, and Michael Götze. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 2137–2142, Marrakech, Morocco, 2008.
- Yuping Zhou and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 69–77, Jeju, South Korea, 2012.