Frontiers in Corpus Annotations II

Proceedings of the Workshop

ACL 2005
Ann Arbor, Michigan
June 29, 2005

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Preface

This volume contains the papers prepared for and presented at the *Frontiers in Corpus Annotation II: Pie in the Sky* workshop held on June 29, 2005 at the University of Michigan, as part of the 2005 annual meeting of the Association of Computational Linguistics.

This workshop is the second in a series of workshops about innovation in corpus annotation and its effect on the future of Computational Linguistics.

I wish to thank and acknowledge the program committee, participants in the workshop, those who contributed and submitted papers, those who participated in the “Pie in the Sky” email list and my wife Jenny for the cover illustration.

Adam Meyers
Computer Science Dept.
New York University

Chair: Adam Meyers, New York University

Program Committee:

Charles J. Fillmore, International Computer Science Institute, Berkeley
Eva Hajicova, Center for Computational Linguistics, Charles University, Prague
Boyan A. Onyshkevych, U.S. Dept. of Defense
Martha Palmer, University of Pennsylvania, Philadelphia
David Farwell, Computing Research Laboratory, New Mexico State University, Las Cruces, NM
Sergei Nirenburg, University of Maryland, Baltimore County
Owen Rambow, Columbia University, NYC
Beth Sundheim, SPAWAR Systems Center, San Diego
Gerald Penn, University of Toronto, Toronto
James Pustejovsky, Brandeis University, Waltham, Mass.

Additional Reviewer: Chadwick McHenry, MITRE

Cover Illustration: Jennifer Eve Meyers

Conference Website: http://nlp.cs.nyu.edu/meyers/frontiers/2005.html

Pie in the Sky Website: http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html

Table of Contents

Conference Program

Wednesday, June 29, 2005, 9:00–18:00

9:00–9:30 *Introduction to Frontiers in Corpus Annotation II: Pie in the Sky*
Adam Meyers

9:30–10:00 *Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference*
James Pustejovsky, Adam Meyers, Martha Palmer and Massimo Poesio

10:00–10:30 *A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations*
Erhard W. Hinrichs, Sandra Kübler and Karin Naumann

10:30–11:00 Break

11–11:30 *Parallel Entity and Treebank Annotation*
Ann Bies, Seth Kulick and Mark Mandel

11:30–12:00 *Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives*
Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber

12–12:30 Discussion

12:30–14:00 Lunch

14:00–14:30 *Investigating the Characteristics of Causal Relations in Japanese Text*
Takashi Inui and Manabu Okumura

14:30–15:00 *A Framework for Annotating Information Structure in Discourse*
Sasha Calhoun, Malvina Nissim, Mark Steedman and Jason Brenier

15:00–15:30 *Annotating Attributions and Private States*
Theresa Wilson and Janyce Wiebe

15:30–16:00 Break

16:00–16:30 *A Parallel Proposition Bank II for Chinese and English*
Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen and Benjamin Snyder

Wednesday, June 29, 2005, 9:00–18:00 (continued)

16:30–17:00 *Semantically Rich Human-Aided Machine Annotation*
Marjorie McShane, Sergei Nirenburg, Stephen Beale and Thomas O'Hara

17:00–17:30 *The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account*
Massimo Poesio and Ron Artstein

17:30–18:00 Discussion

Introduction to Frontiers in Corpus Annotation II Pie in the Sky

Adam Meyers
New York University
meyers@cs.nyu.edu

1 Introduction

The *Frontiers in Corpus Annotation* workshops are opportunities to discuss the state of the art of corpus annotation in computational linguistics. Corpus annotation has pushed the enitre field in new directions by providing new task definitions and new standards of analysis. At the first *Frontiers in Corpus Annotation* workshop at *HLT-NAACL 2004* we compared assumptions underlying different annotation projects in light of both multilingual applications and the pursuit of merged representations that incorporate the result of various annotation projects.

Beginning September, 2004, several researchers have been collaborating to produce detailed semantic annotation of two difficult sentences. The effort aimed to produce a single unified representation that goes beyond what may currently be feasible to annotate consistently or to generate automatically. Rather this "pie in the sky" annotation effort was an attempt at defining a future goal for semantic analysis. We decided to use the "Pie in the Sky" annotation effort (http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html) as a theme for this year's workshop. Consequently this theme has been brought out in many of the papers contained in this volume.

The first 4 papers (Pustejovsky et al., 2005; E. W. Hinrichs and S. Kübler and K. Naumann, 2005; Bies et al., 2005; Dinesh et al., 2005) all discuss some aspect of merging annotation. (Pustejovsky et al., 2005) describes issues that arise for merging argument structures for verbs, nouns and discourse connectives, as well as time and anaphora representations. (E. W. Hinrichs and S. Kübler and K. Naumann, 2005) focuses on the merging of syntactic, morphological, semantic and referential annotation. (E. W. Hinrichs and S. Kübler and K. Naumann, 2005) also points out that the "Pie in the Sky"representation lacks syntactic features. This brings to light an important point of discussion: should linguistic analyses be divided out into separate "levels" corresponding to syntax, morphology, discourse, etc. or should/can a single representation represent all such "levels"? As currently conceived, "Pie in the Sky" is intended to be as "language neutral" as possible – this may make adding a real syntactic level difficult. However, arguably, surface relations can be added on as features to Pie in the Sky, even if we delete or ignore those features for some (e.g., language neutral) purposes. Still, other papers present further difficulties for maintaining a single representation that covers multiple modes of analysis. (Bies et al., 2005) discusses possible conflicts between named entity analyses and syntactic structure and (Dinesh et al., 2005) discusses a conflict between discourse structure and syntactic structure. I think it is reasonable to assume that some such conflicts will be resolvable, e.g., I believe that the named entity conflicts point to shortcomings of the original Penn Treebank analysis. However, the discourse structure/syntactic structure conflicts may be harder to solve. In fact, some annotation projects, e.g., the Prague Dependency Treebank (Hajiĉová and Ceplová, 2000), assume that multiple analyses or "levels" are necessary to describe the full range of phenomena.

The 5th through 7th papers (Inui and Okumura, 2005; Calhoun et al., 2005; Wilson and Wiebe, 2005) investigate some additional types of annotation that were not part of the distributed version of Pie in the Sky, but which could be added in principle. In fact, with help from the authors of (Calhoun et al., 2005), I did incorporate their analysis into the latest version (number 6) of the"Pie in the Sky" annotation. Furthermore, it turns out that some units of Information Structure cross the boundaries of the syntactic/semantic constituents, thus raising the sort of difficulties discussed in the previous paragraph. Specifically, information structure divides sentences into themes and rhemes. For the sample two sentences, the rheme boundaries do correspond to syntactic units, but the theme boundaries cross syntactic boundaries, forming units made up of parts of multiple syntactic constituents.

(Palmer et al., 2005; Xue, 2005) (the eighth and

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 1–4,
Ann Arbor, June 2005.

eleventh papers) make comparisons of annotated phenomena across English and Chinese. It should be pointed out that seven of the papers at this workshop are predominantly about the annotation of English, one is about German annotation and one is about Japanese annotation. These two are the only papers at the workshop that explicitly discuss attempts to apply the same annotation scheme across two languages.

(McShane et al., 2005; Poesio and Artstein, 2005) (the ninth and tenth papers) both pertain to issues about improving the annotation process. (Poesio and Artstein, 2005) discusses some better ways of assessing inter-annotator agreement, particularly when there is a gray area between correct and incorrect annotation. (McShane et al., 2005) discusses the issue of human-aided annotation (human correction of a machine-generated analysis) as it pertains to a single-integrated annotation scheme, similar in many ways to "Pie in the Sky", although it has been in existence for a lot longer.

2 Issues for Discussion

These papers raise a number of important issues for discussion, some of which I have already touched on.

Question 1: Should the community annotate lots of individual phenomena independently of one another or should we assume an underlying framework and perform all annotation tasks so they are compatible with that framework?

Some of the work presented describes the annotation of fairly narrow linguistic phenomena. Pie in the Sky can be viewed as a framework for unifying these annotation schemata into a single representation (a Unified Linguistic Annotation framework in the sense of (Pustejovsky et al., 2005)). Other work presented assumes that the integrated framework is the object of the annotation rather than the result of merging annotations (E. W. Hinrichs and S. Kübler and K. Naumann, 2005; McShane et al., 2005). There are pros and cons to both approaches.

When researchers decide to annotate one small piece of linguistic analysis (verb argument structure, noun argument structure, coreference, discourse structure, etc.), this has the following potential advantages: (1) exploring one phenomenon in depth may provide a better characterization of that phenomenon. If individual phenomena are examined with this level of care, perhaps we will end up with a better overall analysis; (2) a very focused task definition for the annotator may improve interannotator agreement; and (3) it is sometimes easier to analyze a phenomenon in isolation, especially if there is not a large literature of previous work about it – indeed, trying to integrate this new phenomenon before adequately understanding it may unduly bias one's research. However, by ignoring a more complete theory, these annotation projects run the risk of task-based biases, e.g., classifying predication as coreference or coreference as argument-hood. While an underlying all-inclusive theory could be a useful roadmap, unifying the results of several annotation efforts (and resolving inconsistencies) may yield the same result (as suggested in (Pustejovsky et al., 2005)) while maintaining the advantages of investigating the phenomena separately. On the other hand, as this merging process has not come to completion yet, the jury is still out.

Let's say that, for the sake of argument, the reader accepts the research program where individual annotation efforts are slowly merged into one "Pie in the Sky" type system. There is still another obvious question that arises:

Question 2: Why make up a brand new system like "Pie in the Sky" when there are so many existing frameworks around? For example, Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994) assumes a fairly large feature structure that would seem to accommodate every possible level of linguistic analysis (although in practice most authors in that framework only work on the syntactic and semantic portion of that feature structure).

Our initial motivation for starting fresh is that we wanted the framework to use the minimal features necessary to represent the input annotation systems and to extend them as much as possible. In addition, part of the experiment was an aim to keep features in a somewhat language-neutral form and it is not clear that there are existing frameworks that both share this bias and are sufficiently expressive for our purposes. However, ultimately it might be beneficial to convert "Pie in the Sky" to one or more pre-existing frameworks.

So far, we have limited the scope of "Pie in the Sky" to semantic and (recently) some discourse information as well. However, there are some cases where we found it necessary to include syntactic information, e.g., although heads are semantic arguments of adjective modifiers, the surface relation between the head of the noun phrase and its constituents is important for determining other parts of meaning. For example, although *explosive* would bear the same argument relation to *powerful* in both (a) *The explosive is powerful* and (b) *the powerful explosive*, the interpretation of (b) requires that *powerful* be part of the same unit as *explosive*, e.g., for the proper interpretation of *He bought a powerful explosive*. Thus it may seem like a good idea to ultimately fill out "Pie in the Sky" into a larger framework. However, we would still want to be able to pick out the language-neutral components of the analysis from the language-specific ones.

Question 3: D. Farwell, a member of the workshop committee, has pointed out that there are levels within semantics. The question is how should these multiple levels be handled? The annotated examples did not include phenomena such as metaphor, metonymy or idiomaticity that may have multiple interpretations: literal and intended.

For example, an adequate interpretation of *I love listening to Mozart* would require *Mozart* to be decomposed into *music by Mozart* (although arguably the representation of some of the complex discourse references were of this flavor).

3 What's in the Latest Pie in the Sky Analysis

As of this writing, the latest "Pie in the Sky" analysis includes: (1) argument structure of all parts of speech (verbs, nouns, adjectives, determiners, conjunctions, etc.) using the PropBank/NomBank/Discourse Treebank argument labels (ARG0, ARG1, ARG2, ...), reminiscent of Relational Grammar of the 1970s and 1980s (Perlmutter, 1984), (2) some more specifically labeled FrameNet (Baker et al., 1998) roles for these same constituents; (3) morphological and part of speech features; (4) pointers to gazetteers, both real and hypothetical (thanks to B. Sundheim); (5) Veracity/According-To features based on NYU's proposed FactBank annotation scheme; (6) various coreference features including some based on a proposed extension to NomBank; (7) temporal features based on Timex2 (Ferro et al., 2002) and TimeML (Pustejovsky et al., 2004); and (8) Information Structure features based on (Calhoun et al., 2005). For more detail, please see: http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html

4 The Future of "Pie in the Sky"

After this workshop, we plan to retire the current two "Pie in the Sky" sentences and start again with some new text. I observed the following obstacles during this experiment: (1) annotation projects were somewhat hesitant to volunteer their time (so we are extremely grateful to all projects that did so.); (2) the target material was not long enough for some annotation approaches to be able to really make their mark, e.g., two sentences are not so interesting for discourse purposes.; and (3) partially due to its length, some interesting phenomena were not well-represented (idioms, metonymy, etc.)

The lack of volunteers may, in part, be related to the scale of the project. We built the project up slowly and invited people to join in, rather than posting a request for annotations to an international list. Initially, this was necessary just to make the project possible to manage. Additionally, inadequacies of the data were probably barriers for projects that focused on discourse phenomena or phenomena that was not well-represented by our data. Nevertheless, using more data may place too heavy a burden on annotation projects and this could make projects hesitant to participate.

With these issues in mind, I note that several sites annotated two longer documents for the recent U.S. Government sponsored Semantic Annotation Planning Meeting at the University of Maryland. This success was, in part, due to the chance for annotation sites to attract government interest in funding their projects. While we will not attempt to duplicate this workshop, I believe that there is an underlying issue that is very important. The field really needs a single test corpus for all new annotation projects.

This test corpus would meet a number of important needs of the annotation community: (1) it would provide a testbed for new annotation schemata; (2) it would provide a large corpus that is annotated in a fairly complete framework – this way focused annotation projects may be able to more easily write specifications in light of where their particular set of phenomena fit into some larger framework; and (3) it would provide a steady flow of input annotation in order to produce a single unified annotation framework.

To make this idea a reality, we need to obtain a consensus on what people would like to annotate. Additionally, we need volunteers to translate this same corpus into other languages, as we would inevitably choose an English corpus. Of course, if we could find a suitable text that was already translated in multiple languages, this would save time. The perfect text would be article length (loosely defined); include difficult to handle phenomena (idioms, metonymy, etc.); include a wide range of annotatable linguistic phenomena and not have copyright restrictions which would hamper the project. It would, of course, be helpful if the annotation community would provide input on which text to choose – this would avoid a situation where one could not annotate the test text because the target phenomenon is not represented there.

In summary, I have used this introduction to both summarize how the papers of this workshop fit together, to propose some unifying themes for discussion, and to propose an agenda for how to proceed after the workshop is over. We hope to see some of these ideas come to fruition before "Frontiers in Corpus Annotation III."

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of Coling-ACL98: The 17th International Conference on Computational Linguistics and the 36th Meeting of the Association for Computational Linguistics*, pages 86–90.

A. Bies, S. Kulick, and M. Mandel. 2005. Parallel Entity and Treebank Annotation. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

S. Calhoun, M. Nissim, M. Steedman, and J. Brenier. 2005. A Framework for Annotating Information Struc-

ture in Discourse. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

N. Dinesh, A. Lee, E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

E. W. Hinrichs and S. Kübler and K. Naumann. 2005. A Unified Reprewsentation for Morphological, Syntactic, Semantic, and Referential Annotations. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2002. Instruction Manual for the Annotation of Temporal Expressions. MITRE Washington C3 Center, McLean, Virginia.

Eva Hajiĉová and Mark'eta Ceplová. 2000. Deletions and Their Reconstruction in Tectogrammatical Syntactic Tagging of Very Large Corpora. In *Proceedings of Coling 2000: The 18th International Conference on Computational Linguistics*, pages 278–284.

T. Inui and M. Okumura. 2005. Investigating the Characteristics of Causal Relations in Japanese Text. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

M. McShane, S. Nirenburg, S. Beale, and T. O'Hara. 2005. Semantically Rich Human-Aided Machine Annotation. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

M. Palmer, N. Xue, O. Babko-Malaya, J. Chen, and B. Snyder. 2005. A Parallel Proposition Bank II for Chinese and English. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

David. M. Perlmutter. 1984. *Studies in Relational Grammar 1*. The University of Chicago Press, Chicago.

M. Poesio and R. Artstein. 2005. The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago and Stanford.

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2004. The Specification Language TimeML. In I. Mani, J. Pustejovsky, and R. Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press, Oxford.

J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

T. Wilson and J. Wiebe. 2005. Annotating Attributions and Private States. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

N. Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference

James Pustejovsky, Adam Meyers, Martha Palmer, Massimo Poesio

Abstract

Many recent annotation efforts for English have focused on pieces of the larger problem of semantic annotation, rather than initially producing a single unified representation. This paper discusses the issues involved in merging four of these efforts into a unified linguistic structure: PropBank, NomBank, the Discourse Treebank and Coreference Annotation undertaken at the University of Essex. We discuss resolving overlapping and conflicting annotation as well as how the various annotation schemes can reinforce each other to produce a representation that is greater than the sum of its parts.

1. Introduction

The creation of the Penn Treebank (Marcus et al, 1993) and the word sense-annotated SEMCOR (Fellbaum, 1997) have shown how even limited amounts of annotated data can result in major improvements in complex natural language understanding systems. These annotated corpora have led to high-level improvements for parsing and word sense disambiguation (WSD), on the same scale as previously occurred for Part of Speech tagging by the annotation of the Brown corpus and, more recently, the British National Corpus (BNC) (Burnard, 2000). However, the creation of semantically annotated corpora has lagged dramatically behind the creation of other linguistic resources: in part due to the perceived cost, in part due to an assumed lack of theoretical agreement on basic semantic judgments, in part, finally, due to the understandable unwillingness of research groups to get involved in such an undertaking. As a result, the need for such resources has become urgent.

Many recent annotation efforts for English have focused on pieces of the larger problem of semantic annotation, rather than producing a single unified representation like Head-driven Phrase Structure Grammar (Pollard and Sag 1994) or the Prague Dependency Tectogramatical Representation (Hajicova & Kucerova, 2002). PropBank (Palmer et al, 2005) annotates predicate argument structure anchored by verbs. NomBank (Meyers, et. al., 2004a) annotates predicate argument structure anchored by nouns. TimeBank (Pustejovsky et al, 2003) annotates the temporal features of propositions and the temporal relations between propositions. The Penn Discourse Treebank (Miltsakaki et al 2004a/b) treats discourse connectives as predicates and the sentences being joined as arguments. Researchers at Essex were responsible for the coreference markup scheme developed in MATE (Poesio et al, 1999; Poesio, 2004a) and have annotated corpora using this scheme including a subset of the Penn Treebank (Poesio and Vieira, 1998), and the GNOME corpus (Poesio, 2004a). This paper discusses the issues involved in creating a *Unified Linguistic Annotation (ULA)* by merging annotation of examples using the schemata from these efforts. Crucially, all individual annotations can be kept separate in order to make it easy to produce alternative annotations of a specific type of semantic information without need to modify the annotation at the other levels. Embarking on separate annotation efforts has the advantage of allowing researchers to focus on the difficult issues in each area of semantic annotation and the disadvantage of inducing a certain amount of tunnel vision or task-centricity – annotators working on a narrow task tend to see all phenomena in light of the task they are working on, ignoring other factors. However, merging these annotation efforts allows these biases to be dealt with. The result, we believe, could be a more detailed semantic account than possible if the ULA had been the initial annotation effort rather than the result of merging.

There is a growing community consensus that general annotation, relying on linguistic cues, and in particular lexical cues, will produce an enduring resource that is useful, replicable and portable. We provide the beginnings of one such level derived from several distinct annotation efforts. This level could provide the foundation for a major advance in our ability to automatically extract salient relationships from text. This will in turn facilitate breakthroughs in message understanding, machine translation, fact retrieval, and information retrieval.

2. The Component Annotation Schemata

We describe below existing independent annotation efforts, each one of which is focused on a specific aspect of the semantic representation task: semantic role labeling,

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 5–12,
Ann Arbor, June 2005.

coreference, discourse relations, temporal relations, etc. They have reached a level of maturity that warrants a concerted attempt to merge them into a single, unified representation, ULA. There are several technical and theoretical issues that will need to be resolved in order to bring these different layers together seamlessly. Most of these approaches have annotated the same type of data, Wall Street Journal text, so it is also important to demonstrate that the annotation can be extended to other genres such as spoken language. The demonstration of success for the extensions would be the training of accurate statistical semantic taggers.

PropBank: The Penn Proposition Bank focuses on the argument structure of verbs, and provides a corpus annotated with semantic roles, including participants traditionally viewed as arguments and adjuncts. An important goal is to provide consistent semantic role labels across different syntactic realizations of the same verb, as in *the window* in *[ARG0 John] broke [ARG1 the window]* and *[ARG1 The window] broke*. Arg0 and Arg1 are used rather than the more traditional Agent and Patient to keep the annotation as theory-neutral as possible, and to facilitate mapping to richer representations. The 1M word Penn Treebank II Wall Street Journal corpus has been successfully annotated with semantic argument structures for verbs and is now available via the Penn Linguistic Data Consortium as PropBank I (Palmer, et. al., 2005). Coarse-grained sense tags, based on groupings of WordNet senses, are being added, as well as links from the argument labels in the Frames Files to FrameNet frame elements. There are close parallels to other semantic role labeling projects, such as FrameNet (Baker, et. al., 1998; Fillmore & Atkins, 1998; Fillmore & Baker, 2001), Salsa (Ellsworth, et.al, 2004), Prague Tectogrammatics (Hajicova & Kucerova, 2002) and IAMTC, (Helmreich, et. al., 2004)

NomBank: The NYU NomBank project can be considered part of the larger PropBank effort and is designed to provide argument structure for instances of about 5000 common nouns in the Penn Treebank II corpus (Meyers, et. al., 2004a). PropBank argument types and related verb Frames Files are used to provide a commonality of annotation. This enables the development of systems that can recognize regularizations of lexically and syntactically related sentence structures, whether they occur as verb phrases or noun phrases. For example, given an IE system tuned to a *hiring* scenario (MUC-6, 1995), NomBank and PropBank annotation facilitate generalization over patterns. PropBank and NomBank would both support a single IE pattern stating that the object (ARG1) of *appoint* is *John* and the subject (ARG0) is *IBM*, allowing a system to detect that *IBM hired John* from each of the following strings: *IBM appointed John*, *John was appointed by IBM*, *IBM's appointment of John*, *the appointment of John by IBM* and *John is the current IBM appointee*.

Coreference: Coreference involves the detection of subsequent mentions of invoked entities, as in *George Bush,… he….* Researchers at Essex (UK) were responsible for the coreference markup scheme developed in MATE (Poesio et al, 1999; Poesio, 2004a), partially implemented in the annotation tool MMAX and now proposed as an ISO standard; and have been responsible for the creation of two small, but commonly used anaphorically annotated corpora – the Vieira / Poesio subset of the Penn Treebank (Poesio and Vieira, 1998), and the GNOME corpus (Poesio, 2004a). Parallel coreference annotation efforts funded by ACE have resulted in similar guidelines, exemplified by BBN's recent annotation of Named Entities, common nouns and pronouns. These two approaches provide a suitable springboard for an attempt at achieving a community consensus on coreference.

Discourse Treebank: The Penn Discourse Treebank (PDTB) (Miltsakaki et al 2004a/b) is based on the idea that discourse connectives are predicates with associated argument structure (for details see (Miltsakaki et al 2004a, Miltsakaki et al 2004b). The long-range goal is to develop a large scale and reliably annotated corpus that will encode coherence relations associated with discourse connectives, including their argument structure and anaphoric links, thus exposing a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with discourse connectives. This annotation references the Penn Treebank annotations as well as PropBank, and currently only considers Wall Street Journal text.

TimeBank: The Brandeis TimeBank corpus, funded by ARDA, focuses on the annotation of all major aspects in natural language text associated with temporal and event information (Day, et al, 2003, Pustejovsky, et al, 2004). Specifically, this involves three areas of the annotation: temporal expressions, event-denoting

expressions, and the links that express either an anchoring of an event to a time or an ordering of one event relative to another. Identifying events and their temporal anchorings is a critical aspect of reasoning, and without a robust ability to identify and extract events and their temporal anchoring from a text, the real aboutness of the article can be missed. The core of TimeBank is a set of 200 news reports documents, consisting of WSJ, DUC, and ACE articles, each annotated to TimeML 1.2 specification. It is currently being extended to AQUAINT articles. The corpus is available from the timeml.org website.

3. Unifying Linguistic Annotations

Since September, 2004, researchers representing several different sites and annotation projects have begun collaborating to produce a detailed semantic annotation of two difficult sentences. These researchers aim to produce a single unified representation with some consensus from the NLP community. This effort has given rise to both a listserv email list and this workshop: http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html, http://nlp.cs.nyu.edu/meyers/frontiers/2005.html The merging operations discussed here would seem crucial to the furthering of this effort.

3.1 The Initial *Pie in the Sky* Example

The following two consecutive sentences have been annotated for Pie in the Sky.

Two Sentences From ACE Corpus File NBC20001019.1830.0181

- *but Yemen's president says the FBI has told him the explosive material could only have come from the U.S., Israel or two Arab countries.*
- *and to a former federal bomb investigator, that description suggests a powerful military-style plastic explosive c-4 that can be cut or molded into different shapes.*

Although the full *Pie-in-the-Sky* analysis includes information from many different annotation projects, the Dependency Structure in Figure 1 includes only those components that relate to PropBank, NomBank, Discourse annotation, coreference and TimeBank. Several parts of this representation require further explanation. Most of these are signified by the special arcs, arc labels, and nodes. Dashed lines represent transparent arcs, such as the transparent dependency between the argument (ARG1) of modal *can* and the *or*. *Or* is transparent in that it allows this dependency to pass through it to *cut* and *mold*. There are two small arc loops -- *investigator* is its own ARG0 and *description* is its own ARG1. *Investigator* is a relational noun in NomBank. There is assumed to be an underlying relation between the *Investigator* (ARG0), the beneficiary or employer (the ARG2) and the item investigated (ARG1). Similarly, *description* acts as its own ARG1 (the thing described). There are four special coreference arc labels: ARG0-CF, ARG-ANAPH, EVENT-ANAPH and ARG1-SBJ-CF. At the target of these arcs are pointers referring to phrases from the previous sentence or previous discourse. The first three of these labels are on arcs with the noun *description* as their source. The ARG0-CF label indicates that the phrase *Yemen's president* (**1**) is the ARG0, the one who is doing the describing. The EVENT-ANAPH label points to a previous mention of the describing event, namely the clause: *The FBI told him the explosive material…* (**3**). However, as noted above, the NP headed by *description* represents the thing described in addition to the action. The ARG-ANAPH label points to the thing that the FBI told him *the explosive material can only come from …* (**2**). The ARG1-SBJ-CF label links the NP from the discourse *what the bomb was made from* as the subject with the NP headed by *explosive* as its predicate, much the same as it would in a copular construction such as: *What the bomb was made from is the explosive C-4*. Similarly, the arc ARG1-APP marks *C-4 as an* apposite, also predicated to the NP headed by *explosive*. Finally, the thick arcs labeled SLINK-MOD represent TimeML SLINK relations between eventuality variables, i.e., the *cut* and *molded* events are modally subordinate to the *suggests* proposition. The merged representation aims to be compatible with the projects from which it derives, each of which analyzes a different aspect of linguistic analysis. Indeed most of the dependency labels are based on the annotation schemes of those projects.

We have also provided the individual PropBank, NomBank and TimeBank annotations below in textual form, in order to highlight potential points of interaction.

PropBank: and [$_{Arg2}$ to a former federal bomb investigator], [$_{Arg0}$ that description] [$_{Rel\_suggest.01}$ suggests] [$_{Arg1}$ [$_{Arg1}$ a powerful military-style plastic explosive c-4] that

[ArgM-MOD can] be [Rel_cut.01 cut] or [Rel_mold.01 molded] [ArgM-RESULT into different shapes]].

NomBank*:* and to a former [Arg2 federal] [Arg1 bomb] [Rel investigator], that description suggests a powerful [Arg2 military] - [Rel style] plastic [Arg1 explosive] c-4 that can be cut or molded into different shapes.

TimeML: and to a former federal bomb investigator, that description [Event = ei1 suggests] a powerful military-style plastic explosive c-4 that can be [Event = ei2 modal='can' cut] or [Event = ei3 modal='can' molded] into different shapes. <SLINK eventInstanceID = ei1 subordinatedEventID = ei2 relType = 'Modal'/> <SLINK eventInstanceID = ei1 subordinatedEventID = ei3 relType = 'Modal'/>

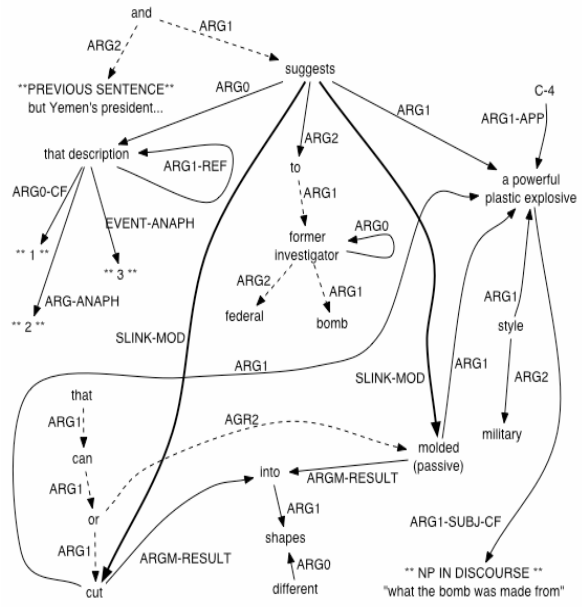


Figure 1. Dependency Analysis of Sentence 2

Note that the subordinating Events indicated by the TimeML SLINKS refer to the predicate argument structures labeled by PropBank, and that the ArgM-MODal also labeled by PropBank contains modality information also crucial to the SLINKS. While the grammatical modal on *cut* and *mold* is captured as an attribute value on the event tag, the governing event predicate *suggest* introduces a modal subordination to its internal argument, along with its relative clause. While this markup is possible in TimeML, it is difficult to standardize (or automate, algorithmically) since arguments are not marked up unless they are event denoting.

3.2 A More Complex Example

To better illustrate the interaction between annotation levels, and the importance of merging information resident in one level but not necessarily in another, consider the sentence below which has more complex temporal properties than the *Pie-in-the-Sky* sentences and its dependency analysis (Figure 2).

According to reports, sea trials for a patrol boat developed by Kazakhstan are being conducted and the formal launch is planned for the beginning of April this year.

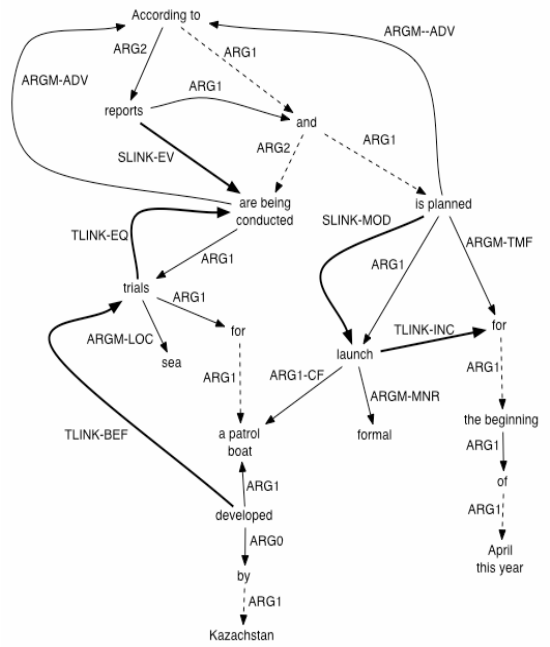


Figure 2. Dependency Analysis of a Sentence with Interesting Temporal Properties

The graph above incorporates these distinct annotations into a merged representation, much like the previous analysis. This sentence has more TimeML annotation than the previous sentence. Note the loops of arcs which show that *According to* plays two roles in the sentence: (1) it heads a constituent that is the ARGM-ADV of the verbs *conducted* and *planned;* (2) it indicates that the information in this entire sentence is attributed to *the reports*. This loop is problematic in some sense because the adverbial appears to modify a constituent that includes itself. In actuality, however, one would expect that the ARGM-ADV role modifies the sentence minus the adverbial, the constituent that you would get if you ignore the transparent arc from ARGM-

ADV to the rest of the sentence. Alternatively, a merging decision may elect to delete the ARGM-ADV arcs, once the more specific predicate argument structure of the sentence adverbial annotation is available.

The **PropBank** annotation for this sentence would label arguments for *develop, conduct* and *plan*, as given below.

[ArgM-ADV According to reports], [Arg1sea trials for *[Arg1 a patrol boat] [Rel_develop.02 developed] [Arg0 by Kazakhstan*]] are being
[Rel_conduct.01 conducted] and [Arg1 the formal launch] is [Rel_plan.01 planned]
[ArgM-TMP for the beginning of April this year].

NomBank would add arguments for *report, trial*, *launch* and *beginning* as follows:

According to [Rel_report.01 reports], [Arg1 [ArgM-LOC sea [Rel_trial.01 trials] [Arg1 for [Arg1-CF_launch.01 a patrol boat] developed by Kazakhstan] are being conducted and the [ArgM-MNR formal] [Rel_launch.01 launch] is planned for the [[REL_beginning.01 beginning] [ARG1 of April this year]].

TimeML, however, focuses on the anchoring of events to explicit temporal expressions (or document creation dates) through TLINKs, as well as subordinating relations, such as those introduced by modals, intensional predicates, and other event-selecting predicates, through SLINKs. For discussion, only part of the complete annotation is shown below.

According to [Event = ei1 reports], sea [Event = ei3 trials] for a boat [Event = ei4 developed] by Kazakhstan are being [Event = ei5 conducted] and the formal [Event = ei6 launch]
is [Event = ei7 planned] for the [Timex3= t1 beginning of April] [Timex3= t2 this year].
<SLINK eventID="ei1" subordinatedEvent="ei5, ei7" relType=EVIDENTIAL/>
<TLINK eventID="ei4" relatedToEvent ="ei3" relType=BEFORE/>
<TLINK eventID="ei6" relatedToTime="t1" relType=IS_INCLUDED />
<SLINK eventID="ei7" subordinatedEvent="ei6" relType="MODAL"/>
<TLINK eventID="ei5" relatedToEvent="ei3" relType=IDENTITY/>

Predicates such as *plan* and nominals such as *report* are lexically encoded to introduce SLINKs with a specific semantic relation, in this case, a "MODAL" relType,. This effectively introduces an intensional context over the subordinated events.

These examples illustrate the type of semantic representation we are trying to achieve. It is clear that our various layers already capture many of the intended relationships, but they do not do so in a unified, coherent fashion. Our goal is to develop both a framework and a process for annotation that allows the individual pieces to be automatically assembled into a coherent whole.

4.0 Merging Annotations

4.1 First Order Merging of Annotation

We begin by discussing issues that arise in defining a single format for a merged representation of PropBank, NomBank and Coreference, the core predicate argument structures and referents for the arguments. One possible representation format would be to convert each annotation into features and values to be added to a larger feature structure. [1] The resulting feature structure would combine stand alone and offset annotation – it would include actual words and features from the text as well as special features that point to the actual text (character offsets) and, perhaps, syntactic trees (offsets along the lines of PropBank/NomBank). Alternative global annotation schemes include annotation graphs (Cieri & Bird, 2001), and MATE (Carletta, et. al., 1999). There are many areas in which the boundaries between these annotations have not been clearly defined, such as the treatment of support constructions and light verbs, as discussed below. Determining the most suitable format for the merged representation should be a top priority.

4.2 Resolving Annotation Overlap

There are many possible interactions between different types of annotation: aspectual verbs have argument labels in PropBank, but are also important roles for temporal relations. Support

[1] The Feature Structure has many advantages as a target representation including: (1) it is easy to add lots of detailed features; and (2) the mathematical properties of Feature Structures are well understood, i.e., there are well-defined rule-writing languages, subsumption and unification relations, etc. defined for Feature Structures (Carpenter, 1992) The downside is that a very informative Feature Structure is difficult for a human to read.

constructions also have argument labels, and the question arises as to whether these should be associated with the support verb or the predicative nominal. Given the sentence *They gave the chefs a standing ovation*, a PropBank component will assign role labels to arguments of *give*; a NomBank component will assign argument structure to *ovation* that labels the same participants. If the representations are equivalent, the question arises as to which of them (or both) should be included in the merged representation. The following graph (Figure 3) is a combined PropBank and NomBank analysis of this sentence. "They" is the ARG0 of both "give" and "ovation"; "the chefs" is the ARG2 of "give", but the "ARG1" of ovation; "ovation" is the ARG1 of "give" and "give" is a support verb for "ovation". For this case, a reasonable choice might be to preserve the argument structure from both NomBank and PropBank, and to do the same for other predicative nominals that have *give* (or *receive, obtain, request…*) as a support verb, e.g., (*give a kiss/hug/squeeze*, *give a lecture/speech*, *give a promotion*, etc.). For other support constructions, such as *take a walk*, *have a headache* and *make a mistake*, the noun is really the main predicate and it is questionable whether the verbal argument structure carries much information, e.g., there are no selection restrictions between light verbs and their subject (ARG0) -- these are inherited from the noun. Thus *make a mistake* selects a different type of subject than *make a gain*, e.g., people and organizations make mistakes, but stock prices make gains. For these constructions, the merged representation might not need to include the (ARG0) relation between the subject of the sentence and *make*, and future propbanking efforts might do well to ignore the shared arguments of such instances and leave them for NomBank. However, the merged representation would inherit PropBank's annotation of some other light verb features including: negation, e.g., *They did not take a walk;* modality, e.g., *They might take a walk;* and sentence adverbials, e.g., *They probably will take a walk.*

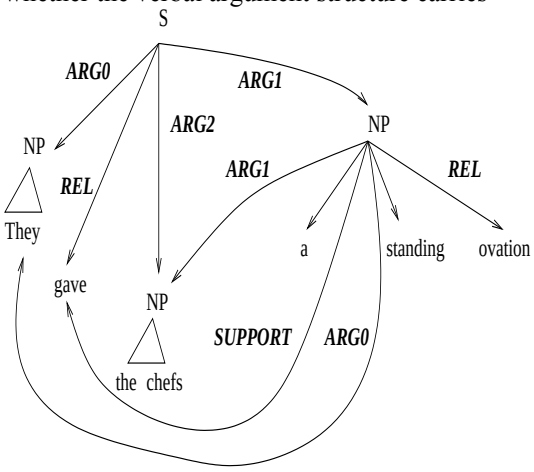


Figure 3. Merged PropBank/NomBank representation of *They gave the chefs a standing ovation.*

4.3 Resolving Annotation Conflicts

Interactions between linguistic phenomena can aid in quality control, and conflicts found during the deliberate merging of different annotations provides an opportunity to correct and fine-tune the original layers. For example, predicate argument structure (PropBank and NomBank) annotation sometimes assumes different constituent structure than the Penn Treebank. We have noticed some tendencies that help resolve these conflicts, e.g., prenominal noun constituents as in *Indianapolis 500,* which forms a single argument in NomBank, is correctly predicted to be a constituent, even though the Penn Treebank II assumes a flatter structure.

Similarly, idioms and multiword expressions often cause problems for both PropBank and NomBank. PropBank annotators tend to view argument structure in terms of verbs and NomBank annotators tend to view argument structure in terms of nouns. Thus many examples that, perhaps, should be viewed as idioms are viewed as special senses of either verbs or nouns. Having idioms detected and marked before propbanking and nombanking could greatly improve efficiency.

Annotation accuracy is often evaluated in terms of inter-annotation consistency. Task definitions may need to err on the side of being more inclusive in order to simplify the annotators task. For example, the NomBank project assumes the following definition of a support verb (Meyers, et.al., 2004b): *"... a verb which takes at least two arguments NP_1 and XP_2 such that XP_2 is an argument of the head of NP_1. For example, in* ***John*** *took a* ***walk****, a support verb (took) shares one of its arguments (John) with the head of its other argument (walk).*" The easiest way to apply this definition is without exception, so it will include idiomatic expressions such as *keep tabs on*, *take place*, *pull strings*. Indeed, the dividing line between support constructions and idioms is difficult to draw (Meyers 2004b). PropBank annotators are also quite comfortable with associating general meanings to the main verbs of idiomatic expressions and labeling their

argument roles, as in cases like *bring home the bacon* and *mince words with.* Since idioms often have interpretations that are metaphorical extensions of their literal meaning, this is not necessarily incorrect. It may be helpful to have the literal dependencies and the idiomatic reading both represented. The fact that both types of meaning are available is evidenced by jokes, irony, and puns.

With respect to idioms and light verbs, TimeML can be viewed as a mediator between PropBank and NomBank. In TimeML, light verbs and the nominalizations accompanying them are marked with two separate EVENT tags. This guarantees an annotation independent of textual linearity and therefore ensures a parallel treatment for different textual configurations. In (a) the light verb construction "make an allusion" is constituted of a verb and an NP headed by an event-denoting noun, whereas in (b) the nominal precedes a VP, which in addition contains a second N:

(a) *Max [made an allusion] to the crime.*

(b) *Several anti-war [demonstrations have taken place] around the globe.*

Both verbal and nominal heads are tagged because they both contribute relevant information to characterizing the nature of the event. The nominal element plays a role in the more semantically based task of event classification. On the other hand, the information in the verbal component is important at two different levels: it provides the grammatical features typically associated with verbal morphology, such as tense and aspect, and at the same time it may help in disambiguating cases like *take/give a class*, *make/take a phone call*. The two tagged events are marked as identical by a TLINK introduced for that purpose. The TimeML annotation for the example in (a) is provided below.

Max [$_{\text{Event = ei1}}$ made] an [$_{\text{Event = ei2}}$ allusion] to the crime.

<TLINK eventID="ei1"relatedToEvent="ei2" relType=IDENTITY>

Some cases of support in NomBank could also be annotated as "bridging" anaphora. Consider the sentence: *The pieces make up the whole*. It is unclear whether *make up* is a support verb linking *whole* as the ARG1 of *pieces* or if *pieces* is linked to *whole* by bridging anaphora. There are also clearer cases. In *Nastase, a rival player defeated Jimmy Connors in the third round,* the word *rival* and *Jimmy Connors* are clearly linked by bridging. However, a wayward NomBank annotator might construct a support chain (*player* + *defeated*) to link *rival* with its ARG1 *Jimmy Connors*. In such a case, a merging of annotation could reveal annotation errors. In contrast, a NomBank annotator would be correct in linking *John* as an argument of *walk* in *John took a series of walks* (the support chain *took* + *series* consists of a support verb and a transparent noun), but this may not be obvious to the non-NomBanker. Thus the merging of annotation may result in the more consistent specifications for all.

In our view, this process of annotating all layers of information and then merging them in a supervised manner, taking note of the conflicts, is a necessary prerequisite to defining more clearly the boundaries between the different types of annotation and determining how they should fit together. Other areas of annotation interaction include: (1) NomBank and Coreference, e.g. deriving that *John teaches Mary* from *John is Mary's teacher* involves: (a) recognizing that *teacher* is an argument nominalization such that the teacher is the ARG0 of *teach* (the one who teaches); and (b) marking *John* and *teacher* as being linked by predication (in this case, an instance of type coreference); and (2) Time and Modality - when a fact used to be true, there are two time components: one in which the fact is true and one in which it is false. Clearly more areas of interaction will emerge as more annotation becomes available and as the merging of annotation proceeds.

5. Summary

We proposed a way of taking advantage of the current practice of separating aspects of semantic analysis of text into small manageable pieces. We propose merging these pieces, initially in a careful, supervised way, and hypothesize that the result could be a more detailed semantic analysis than was previously available. This paper discusses some of the reasons that the merging process should be supervised. We primarily gave examples involving the interaction of PropBank, NomBank and TimeML. However, as the merging process continues, we anticipate other conflicts that will require resolution.

References

C. F. Baker, F. Collin, C. J. Fillmore, and J. B. Lowe (1998), The Berkeley FrameNet project. In *Proc. of COLING/ACL-98*, 86--90

O. Babko-Malaya, M. Palmer, X. Nianwen, S. Kulick, A. Joshi (2004), Propbank II, Delving Deeper, In *Proc. of HLT-NAACL Workshop: Frontiers in Corpus Annotation*.

R. Carpenter (1992), *The Logic of Typed Feature Structures*. Cambridge Univ. Press.

J. Carletta and A. Isard (1999), The MATE Annotation Workbench: User Requirements. In Proc. of the ACL Workshop: Towards Standards and Tools for Discourse Tagging. Univ. of Maryland, 11-17

C. Cieri and S. Bird (2001), Annotation Graphs and Servers and Multi-Modal Resources: Infrastructure for Interdisciplinary Education, Research and Development Proc. of the ACL Workshop on Sharing Tools and Resources for Research and Education, 23-30

D. Day, L. Ferro, R. Gaizauskas, P. Hanks, M. Lazo, J. Pustejovsky, R. Saurí, A. See, A. Setzer, and B. Sundheim (2003), The TIMEBANK Corpus. *Corpus Linguistics*.

M. Ellsworth, K. Erk, P. Kingsbury and S. Pado (2004), PropBank, SALSA, and FrameNet: How Design Determines Product, in *Proc. of LREC 2004 Workshop: Building Lexical Resources from Semantically Annotated Corpora* .

C. Fellbaum (1997), *WordNet: An Electronic Lexical Database*, MIT Press..

C. J. Fillmore and B. T. S. Atkins (1998), FrameNet and lexicographic relevance. In the *Proc. of the First International Conference on Language Resources and Evaluation.*

C. J. Fillmore and C. F. Baker (2001), Frame semantics for text understanding. In *Proc. of NAACL WordNet and Other Lexical Resources Workshop*.

E. Hajivcova and I. Kuvcerov'a (2002). Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. In the *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, 846--851.

S. Helmreich, D. Farwell, B. Dorr, N. Habash, L. Levin, T. Mitamura, F. Reeder, K. Miller, E. Hovy, O. Rambow and A. Siddharthan,(2004), Interlingual Annotation of Multilingual Text Corpora, *Proc. of the HLT-EACL Workshop on Frontiers in Corpus Annotation*.

A, Meyers, R. Reeves, C. Macleod, R, Szekely, V. Zielinska, B. Young, and R. Grishman (2004a), The NomBank Project: An Interim Report, *Proc. of HLT-EACL Workshop: Frontiers in Corpus Annotation*.

A. Meyers, R. Reeves, and C. Macleod (2004b), NP-External Arguments: A Study of Argument Sharing in English. In *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*.

E. Miltsakaki, R. Prasad, A. Joshi and B. Webber. (2004a), The Penn Discourse Treebank. *In Proc. 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

E. Miltsakaki, R. Prasad, A. Joshi and B. Webber (2004b), Annotation of Discourse Connectives and their Arguments, in *Proc. of HLT-NAACL Workshop: Frontiers in Corpus Annotation*

M. Marcus, B. Santorini, and M. Marcinkiewicz (1993), Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313--330.

M. Palmer, D. Gildea, P. Kingsbury (2005), The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal*, *31:1*.

M. Poesio (2004a), The MATE/GNOME Scheme for Anaphoric Annotation, Revisited, *Proc. of SIGDIAL*

M. Poesio (2004b), Discourse Annotation and Semantic Annotation in the GNOME Corpus, *Proc. of ACL Workshop on Discourse Annotation*.

M. Poesio and M. Alexandrov-Kabadjov (2004), A general-purpose, off-the-shelf system for anaphora resol.. *Proc. of LREC*.

M. Poesio, F. Bruneseaux, and L. Romary (1999), The MATE meta-scheme for coreference in dialogues in multiple language, *Proc. of the ACL Workshop on Standards for Discourse Tagging.*

M. Poesio and R. Vieira (1998), A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2).

C. Pollard and I. A. Sag (1994), *Head-driven phrase structure grammar.* Univ. of Chicago Press.

J. Pustejovsky, R. Saurí, J. Castaño, D. R. Radev, R. Gaizauskas, A. Setzer, B. Sundheim and G. Katz (2004), Representing Temporal and Event Knowledge for QA Systems. In Mark T. Maybury (ed.), *New Directions in Question Answering*, MIT Press.

J. Pustejovsky, B. Ingria, R. Saurí, J. Castaño, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani (2003), The Specification Language TimeML. In I. Mani, J. Pustejovsky, and R. Gaizauskas, editors, *The Language of Time: A Reader*. Oxford Univ. Press.

A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations

Erhard W. Hinrichs, Sandra Kübler, Karin Naumann
SfS-CL, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{eh,kuebler,knaumann}@sfs.uni-tuebingen.de

Abstract

This paper reports on the SYN-RA (SYNtax-based Reference Annotation) project, an on-going project of annotating German newspaper texts with referential relations. The project has developed an inventory of anaphoric and coreference relations for German in the context of a unified, XML-based annotation scheme for combining morphological, syntactic, semantic, and anaphoric information. The paper discusses how this unified annotation scheme relates to other formats currently discussed in the literature, in particular the annotation graph model of Bird and Liberman (2001) and the pie-in-the-sky scheme for semantic annotation.

1 Introduction

The purpose of this paper is threefold: (i) it discusses an annotation scheme for referential relations for German that is significantly broader in scope than existing schemes for the same task and language and that also goes beyond the inventory of anaphoric relations included in the pie-in-the-sky sample feature structures[1], (ii) it presents a unified, XML-based annotation scheme for combining morphological, syntactic, semantic, and anaphoric information, and (iii) it discusses how this unified annotation scheme relates to other formats currently discussed in the literature, in particular the annotation graph model of Bird and Liberman (2001) and the pie-in-the-sky scheme for semantic annotation[2].

2 Referential Relations

This section introduces the inventory of referential relations adopted in the SYN-RA project. We define *referential relations* as a cover-term for all contextually dependent reference relations. The inventory of such relations adopted for SYN-RA is inspired by the annotation scheme first developed in the MATE project (Davies et al., 1998). However, it takes a cautious approach in that it only adopts those referential relations from MATE for which the developers of MATE report a sufficiently high level of inter-annotator agreement (Poesio et al., 1999).

SYN-RA currently uses the following subset of relations: *coreferential*, *anaphoric*, *cataphoric*, *bound*, *split antecedent*, *instance*, and *expletive*. The potential markables are definite NPs, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns.

There is a second research effort under way at the European Media Laboratory Heidelberg, which also annotates German text corpora and dialog data with referential relations. Since their corpora are not publicly available, it is difficult to verify their inventory of referential relations. Kouchnir (2003) has used their data and describes the relations *anaphoric*, *coreferential*, *bridging*, and *none*.

Following van Deemter and Kibble (2000), we define a *coreference relation* to hold between two

[1] See e.g. nlp.cs.nyu.edu/meyers/pie-in-the-sky/analysis5.

[2] See nlp.cs.nyu.edu/meyers/pie-in-the-sky/pie-in-the-sky-descript.html.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 13–20,
Ann Arbor, June 2005.

NPs just in case they refer to the same extra-linguistic referent in the real world. In the following example, a coreference relation exists between the noun phrases [1] and [2], and an *anaphoric relation* between the noun phrase [2] and the personal pronoun [3]. Since noun phrases [1] and [2] are coreferential, all three NPs belong to the same coreference chain. In keeping with the MUC-6 annotation standard[3], we establish the anaphoric relations of a pronoun only to its most recently mentioned antecedent.

(1) [1 Der neue Vorsitzende der Gewerkschaft
The new chairman of the union
Erziehung und Wissenschaft] heißt [2 Ulli
Education and Science is called Ulli
Thöne]. [3 Er] wurde gestern mit 217
Thöne. He was yesterday with 217
von 355 Stimmen gewählt.
out of 355 votes elected.

'The new chairman of the union of educators and scholars is called Ulli Thöne. He was elected yesterday with 217 of 355 votes.'

Cataphoric relations hold between a preceding pronoun and a following antecedent within the same sentence, even if this antecedent has already been mentioned within the preceding text. An example for a cataphoric relation is shown in (2).

(2) Vier Wochen sind [sie] nun schon in Berlin,
Four weeks are they now already in Berlin,
[die 220 Albaner aus dem Kosovo].
the 220 Albanians from the Kosovo.

'They have already been in Berlin for four weeks, the 200 Albanians from Kosovo.'

The relation *bound* holds between anaphoric expressions and quantified noun phrases as their antecedents (see example (3)).

(3) [Niemandem] fällt es schwer, das Bild
To nobody is it difficult, the picture
vor [sich] zu sehen.
in front of himself to see.

'Nobody has trouble imagining the picture.'

[3]See www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html.

The *split antecedent relation* holds between coordinate NPs/plural pronouns and pronouns/definite NPs referring to one member of the plural expression. In example (4), the indefinite pronoun *beide* enters into two split antecedent relations, with noun phrases 1 and 2.

(4) Aber plötzlich gibt es da einen völlig
But suddenly gives it there a completely
unglaubwürdig und grotesk wirkenden
implausible and grotesque seeming
Anruf [1 des Detektiven] bei [2 der
phone call of the detective to the
Mutter des Opfers] , [beide] weinen
mother of the victim , both cry
sich minutenlang etwas
themselves for some minutes something
vor , ...
verb part , ...

'But suddenly, there is a completely implausible and grotesque phone call from the detective to the mother of the victim, they both cry at each other for several minutes, ...'

An *instance relation* exists between a preceding/following pronoun and its NP antecedent when the pronoun refers to a particular instantiation of the class identified by the NP.

(5) Die konservativen Kräfte warten ja nur
The conservative powers wait just only
darauf, ihm [Sätze] um die Ohren zu
for that, him sentences around the ears to
hauen wie [jenen von den 16
hit like the one about the 16
Mittelstrecklern], denen er in vier
middle-distance runners, to whom he in four
Wochen die Viererkette
weeks the double full-back formation
beibringe.
teaches.

'The conservative powers are just waiting to bombard him with sentences like the one about the 16 middle-distance runners who he is teaching the double full-back formation in four weeks.'

In sentence (5), the relation between the two bracketed NPs is an example of such an instance relation since the second NP is a particular instantiation of the referent denoted by the first NP.

A third person singular neuter pronoun *es* is marked as *expletive* if it has no proper antecedent. This is the case for presentational *es* in example (6), impersonal passive as in example (7), or *es* as subject for verbs without an agent as in example (8).

(6) [1 Es] zeichnet sich die konkrete Möglichkeit
It emerges the concrete possibility
ab.
verb part.
'The concrete possibility emerges.'

(7) [Es] wird bis zum Morgen getanzt.
There is until the morning danced.
'People are dancing until morning.'

(8) [Es] steht schlecht um ihn.
It stands bad for him.
'He is in a bad way.'

Apart from expletive uses of *es* and anaphoric uses with an NP antecedent, the pronoun *es* can also be used in cases of event anaphora as in sentence (9). Here *es* refers to the event of Jochen's winning the lottery. Currently, the annotation in SYN-RA is restricted to NP anaphora and therefore event anaphors such as in sentence (9) remain unannotated for anaphora.

(9) Jochen hat im Lotto gewonnen. Aber er
Jochen has in the lottery won. But he
weiss es noch nicht.
knows it yet not.
'Jochen has won the lottery. But he does not know it yet.'

The annotation of such relations is performed manually with the annotation tool MMAX (Müller and Strube, 2003). Its graphical user interface allows for easy selection of the relevant markables and the accompanying relation between the contextually dependent expression and its antecedent.

3 Automatic Extraction of Markables and of Semantic Information

Annotation of referential relations involves two main tasks: the identification of markables, i.e., identifying the class of expressions that can enter into referential relations, and the identification of the particular referential relations that two or more expressions enter into. Identification of markables requires at least partial syntactic annotation of the text. If referential relations need to be annotated from plain text, then markables must be identified semi-automatically from the output of a chunker or full parser, if available, or otherwise completely manually. However, in each of these two scenarios, identification of markables is a time-consuming process. In case of semi-automatic annotation, the effort required depends on the quality of the parser, but will require at least some amount of manual post-correction of the parser output.

Identification of markables is considerably easier for treebank data since treebanks already provide the necessary syntactic information. For German, there are currently two large-scale treebanks available: the NEGRA/TIGER (Brants et al., 2002) treebank and the Tübingen treebanks for spoken and written German (Stegmann et al., 2000; Telljohann et al., 2003). All the treebanks were annotated with the help of the annotation tool Annotate (Plaehn, 1998). The treebank annotations are available in the Annotate export format (Brants, 1997) and in an XML format.

The SYN-RA project is based on the Tübingen treebank of written German (TüBa-D/Z). This treebank uses as its data source a collection of articles of the German daily newspaper *taz* (*die tageszeitung*). The treebank currently comprises appr. 15 000 sentences, with a new release of 7 000 additional sentences scheduled for June of this year.

Due to its fine grained syntactic annotation, the TüBa-D/Z treebank data are ideally suited as a basis for the identification of markables and for extracting relevant syntactic and semantic properties for each markable. The TüBa-D/Z annotation scheme distinguishes four levels of syntactic constituency: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word or-

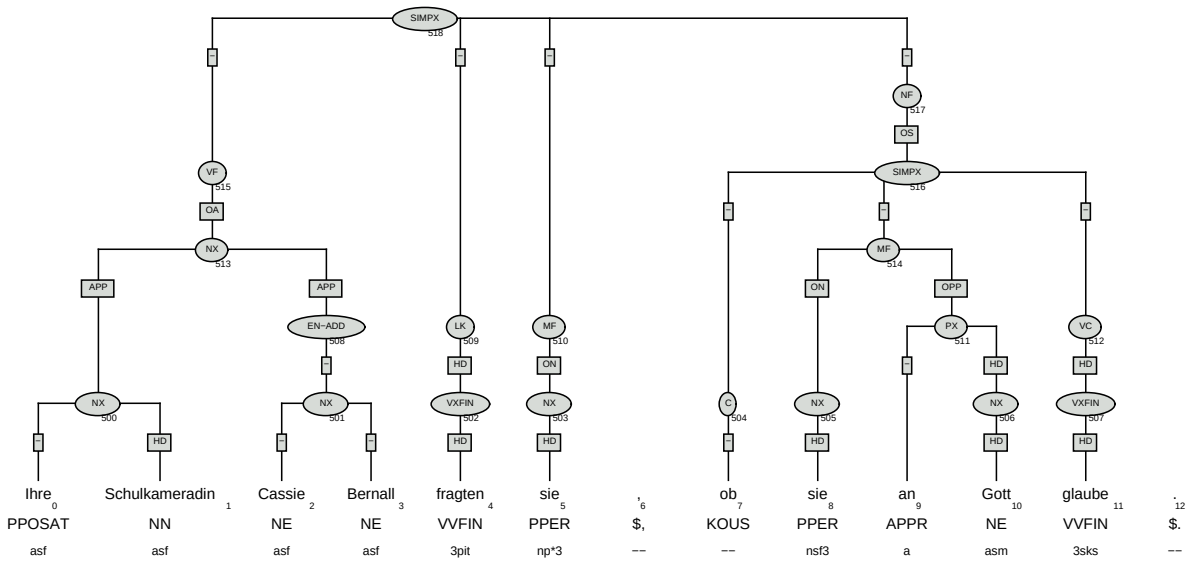


Figure 1: A sample tree from the TüBa/D-Z treebank.

der regularities among different clause types of German and which are widely accepted among descriptive linguists of German (cf. e.g. (Drach, 1937; Höhle, 1986)). The TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question.

Figure 1 shows an example tree from the TüBa-D/Z treebank for sentence (10). The sentence is divided into two clauses (SIMPX), and each clause is subdivided into topological fields. The main clause is made up of the following fields: VF (mnemonic for: *Vorfeld* – 'initial field') contains the sentence-initial, topicalized constituent. LK (for: *linke Satzklammer* – 'left sentence bracket') is occupied by the finite verb. MF (for: *Mittelfeld* – 'middle field') contains adjuncts and complements of the main verb. NF (for: *Nachfeld* – 'final field') contains extraposed material – in this case an indirect yes/no question. The subordinate clause is again divided into three topological fields: C (for: *Komplementierer* – 'complementizer'), MF, and VC (for: *Verbalkomplex* – verbal complex). Edge labels are rendered in boxes and indicate grammatical functions. The sentence-initial NX (for: *noun phrase*) is marked as OA (for: *accusative complement*), the pronouns *sie* in the main and subordinate clause as ON (for: *nominative complement*).

(10) Ihre Schulkameradin Cassie Bernall fragten
Their fellow student Cassie Bernall asked
sie , ob sie an Gott
they[subj] , whether she[subj] in God
glaube.
believes.
'They asked their fellow student Cassie Bernall whether she believes in God.'

Topological field information and grammatical function information is crucial for anaphora resolution since binding-theory constraints crucially rely on sentence-structure (if the binding theory principles are stated configurationally (Chomsky, 1981)) or on argument-obliqueness (if the binding theory principles are stated in terms of argument structure, as in (Pollard and Sag, 1994)). In the case at hand, the subject pronoun of the main clause, *sie*, cannot be anaphorically related to the object NP *Ihre Schulkameradin Cassie Bernall* since they are co-arguments of the same verb. However, the possessive pronoun *ihre* and the subject pronoun *sie* of the subordinate clause, can be and, in fact, are anaphorically related, since they are not co-arguments of the same verb. This can be directly inferred from the treebank annotation, specifically from the sentence structure and the grammatical function information

encoded on the edge labels. Most published computational algorithms of anaphora resolution, including (Hobbs, 1978; Lappin and Leass, 1994; Ingria and Stallard, 1989), rely on such binding-constraint filters to minimize the set of potential antecedents for pronouns and reflexives.

As already pointed out, the sample sentence contains four markables: one possessive pronoun *Ihre*, two occurrences of the pronoun *sie* and one complex NP *Ihre Schulkameradin Cassie Bernall*. The latter NP is a good example of SYN-RA's longest-match principle for identifying markables. In case of complex NPs, the entire NP counts as a markable, but so do its subconstituents – in the case at hand, particularly the possessive pronoun *ihre*. All of this information can be directly derived from the treebank account. Compared to other annotation efforts for German where markables have to be chosen manually (Müller and Strube, 2003), manual annotation in the SYN-RA project can, thus, be restricted to the selection of the appropriate referential relations between referentially dependent expressions and their nominal antecedents.

4 The Unified, XML-based Annotation Scheme

The annotation of referential expressions is embedded in a unified format which also contains morphological, syntactic, and semantic information. The annotation scheme is represented in XML, the widely acknowledged standard for exchanging data, which guarantees portability and re-usability of the data. Each sentence, as well as all words and all nodes in the syntactic structure, are assigned a unique ID. These IDs are used in the annotation of referential relations. The annotation of the treebank sentence 11976 (cf. example (10)) is shown in Figure 2.

The sentence number is encoded as the ID of the sentence. The first word, *Ihre*, has an anaphoric relation to a noun phrase in the previous sentence. This relation is marked in the element *anaphora*, which gives the antecedent as node 517 of sentence 11975, i.e. the previous sentence. The other two anaphoric relations are sentence-internal, the first personal pronoun *sie* having *Ihre* (id: s11976w0) as antecedent, the second one the noun phrase *Ihre Schulfreundin Cassie Bernall* (id: s11976n513). The annotation of the first personal pronoun is an example for the annotation of an anaphoric chain. *Ihre* and *sie* belong to the same chain. However, in order to facilitate the extraction of direct relations, such chains are represented in a way that each anaphoric expression refers to the last occurrence of an antecedent.

The SYN-RA scheme is very similar to the MUC-6 coreference annotation scheme[4] but it is more powerful in two respects: As described above, the inventory is not restricted to coreference and anaphoric relations, it also covers e.g. instance relations or split antecedent relations. The latter relation is also the reason for encoding the relational information as XML elements, and not as attributes of a word or a node. If an anaphor enters into a split antecedent relation, it has more than one distinct antecedent. In this case, the element *anaphora* has two (or more) relations. Such an example is graphically displayed for sentence (4) in Figure 3. The relevant XML representation of the complex entry for the word *beide* is shown in Figure 4.

5 Related Work

This section discusses how the unified SYN-RA annotation scheme relates to other formats currently discussed in the literature, in particular the pie-in-the-sky scheme for semantic annotation[5] and the annotation graph model of (Bird and Liberman, 2001). While these two annotation schemes are by no means the only contenders for corpus annotation standards in the literature, they are certainly among the most ambitious and promising.

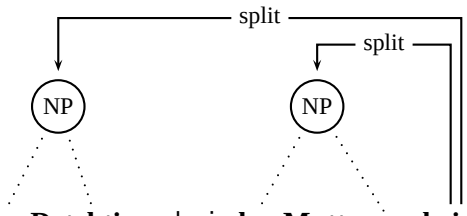
While the pie-in-the-sky scheme is clearly still under development, the following characteristics and goals can already be gleaned from its webpage and the annotation examples presented there: The annotation is feature-structure-based and incorporates various levels of linguistic annotation, in particular a PROPBANK style predicate-argument structure, dependency style syntactic information, as well as morpho-syntactic and word class information. All this information is rooted in the attributes needed for predicate-argument assignment,

[4] See www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html.

[5] See nlp.cs.nyu.edu/meyers/pie-in-the-sky/pie-in-the-sky-descript.html.

```
<sentence  id="s11976">
  <node id="s11976n518" cat="SIMPX" func="--" parent="0">
    <node id="s11976n515" cat="VF" func="-">
      <node id="s11976n513" cat="NX" func="OA">
        <node id="s11976n500" cat="NX" func="APP">
          <word id="s11976w0" form="Ihre" pos="PPOSAT" morph="asf" func="-">
            < anaphora>
              < relation type="ana" antecedent="s11975n517"/>
            </anaphora> </word>
          <word id="s11976w1" form="Schulkameradin" pos="NN" morph="asf" func="HD"/>
        </node>
        <node id="s11976n508" cat="EN-ADD" func="APP">
          <node id="s11976n501" cat="NX" func="-">
            <word id="s11976w2" form="Cassie" pos="NE" morph="asf" func="-"/>
            <word id="s11976w3" form="Bernall" pos="NE" morph="asf" func="-"/>
          </node> </node> </node> </node>
    <node id="s11976n509" cat="LK" func="-">
      <node id="s11976n502" cat="VXFIN" func="HD">
        <word id="s11976w4" form="fragten" pos="VVFIN" morph="3pit" func="HD"/>
      </node> </node>
    <node id="s11976n510" cat="MF" func="-">
      <node id="s11976n503" cat="NX" func="ON">
        <word id="s11976w5" form="sie" pos="PPER" morph="np*3" func="HD">
           < anaphora>
            < relation type="ana" antecedent="s11976w1"/>
          </anaphora> </word> </node> </node>
    <word id="s11976w6" form="," pos="$," morph="--" func="--" parent="0"/>
    <node id="s11976n517" cat="NF" func="-">
      <node id="s11976n516" cat="SIMPX" func="OS">
        <node id="s11976n504" cat="C" func="-">
          <word id="s11976w7" form="ob" pos="KOUS" morph="--" func="-"/>
        </node>
        <node id="s11976n514" cat="MF" func="-">
          <node id="s11976n505" cat="NX" func="ON">
            <word id="s11976w8" form="sie" pos="PPER" morph="nsf3" func="HD">
               < anaphora>
                 < relation type="ana" antecedent="s11976n513"/>
               </anophora> </word> </node>
          <node id="s11976n511" cat="PX" func="OPP" comment="">
              <word id="s11976w9" form="an" pos="APPR" morph="a" func="-"/>
              <node id="s11976n506" cat="NX" func="HD">
                <word id="s11976w10" form="Gott" pos="NE" morph="asm" func="HD"/>
              </node> </node> </node>
        <node id="s11976n512" cat="VC" func="-">
          <node id="s11976n507" cat="VXFIN" func="HD">
            <word id="s11976w11" form="glaube" pos="VVFIN" morph="3sks" func="HD"/>
          </node> </node> </node> </node> </node>
   <word form="." pos="$." morph="--" func="--" parent="0"/>
</sentence>
```

Figure 2: The XML format represents information on all levels of annotation. The words of the sentence and the anaphoric annotation are shown in bold.



Aber plötzlich gibt es da einen ... Anruf **des Detektiven** bei **der Mutter ..., beide** weinen sich minutenlang etwas vor ...

Figure 3: The annotation of the split antecedent relation in sentence (4). For representational reasons, the sentence is shortened and only relevant information is displayed. Syntactic boundaries are shown as dotted lines, anaphoric relations as black lines.

```
<word id="s3426w20" form="beide" pos="PIS" morph="np*" func="HD">
   <anaphora>
      <relation type="split" antecedent="s3426n507"/>
      <relation type="split" antecedent="s3426n526"/>
   </anaphora>
</word>
```

Figure 4: The XML representation of the encoding of split antecedents for the word *beide* in sentence (4). A graphical representation of the relation is shown in Figure 3. The antecedent "s3426n507" refers to the first NP, "s3426n526" to the second one in Figure 3.

with syntactic and morpho-syntactic information distributed among the corresponding elements in the predicate-argument structure representation. Accordingly, semantic representations provide the organizing principle while morpho-syntactic and syntactic information play a subordinated role.

The SYN-RA annotation scheme resembles the pie-in-the-sky scheme in that it also uses one level of representation, in this case hierarchical syntactic structure, as the organizing principle and treats referential relations, grammatical function information, and morpho-syntactic annotation as subordinated types of information. More generally, the pie-in-the-sky and the SYN-RA representations offer a particular view of the annotation, each with its own "perspective": semantics-based (pie-in-the-sky) and syntax-based (SYN-RA).

By contrast, Bird and Liberman's (2001) annotation graphs are intended as a graph-based, multi-layered annotation scheme where each level of linguistic annotation is treated equally, as an independent layer. The graph-based annotation model is powerful enough to also allow groupings of discontinuous constituents and other non-adjacent linguistic phenomena, without having to rearrange the linear order of the input. In both respects, their annotation model is maximally general.

6 Future Directions

In the previous section we have compared two perspective-dependent annotation schemes that use a particular level of linguistic annotation as their primary organizing principle and have contrasted them with the perspective-independent annotation-graph model. We believe that both types of representation models have their independent justification. Perspective-based representations, such as SYN-RA and pie-in-the-sky, are well-justified for particular application scenarios. For example, for text summarization and other semantic tasks, the pie-in-the-sky model seems particularly well-motivated since the pertinent semantic information can be easily extracted from its predicate-argument-structure-rooted feature structures. For other tasks, such as anaphora resolution, for which syntactic information is more relevant, the syntax-based representation of SYN-RA allows for an easier extraction of the relevant information for rule-based, statistical,

and machine-learning approaches to computational anaphora resolution. More generally, perspective-based representations are highly task-dependent. It would be misguided to consider them as ideal, task-independent annotation standards. If one wants to establish a task-independent annotation standard, then a perspective-independent annotation scheme such as the annotation graph model looks like a promising direction for future research. In particular, such research should focus on techniques that allow for easy conversion of perspective-independent representations to task-dependent views of the relevant linguistic information.

References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgaria.

Thorsten Brants, 1997. *The NeGra Export Format for Annotated Corpora*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Sarah Davies, Massimo Poesio, Florence Bruneseaux, and Laurent Romary, 1998. *Annotating Coreference in Dialogues: Proposal for a Scheme for MATE*. MATE.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(2):629–637.

Erich Drach. 1937. *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Robert J. P. Ingria and David Stallard. 1989. A computational mechanism for pronominal reference. In *Proceedings of the 27th Conference of the Association for Computational Linguistics*, pages 262–271, Vancouver, Canada.

Beata Kouchnir. 2003. A machine learning approach to German pronoun resolution. Master's thesis, School of Informatics, University of Edinburgh.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Christoph Müller and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.

Oliver Plaehn, 1998. *Annotate Bedienungsanleitung*. Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany, April.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*, pages 65–74.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press, Chicago, IL.

Rosmary Stegmann, Heike Telljohann, and Erhard W. Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical Report 239, Verbmobil.

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

Parallel Entity and Treebank Annotation

Ann Bies
Linguistic Data Consortium
3600 Market Street, 810
Philadelphia, PA 19104
bies@ldc.upenn.edu

Seth Kulick
Institute for Research
in Cognitive Science
3401 Walnut Street
Suite 400A
Philadelphia, PA 19104
skulick@linc.cis.upenn.edu

Mark Mandel
Linguistic Data Consortium
3600 Market Street, 810
Philadelphia, PA 19104
mamandel@ldc.upenn.edu

Abstract

We describe a parallel annotation approach for PubMed abstracts. It includes both entity/relation annotation and a treebank containing syntactic structure, with a goal of mapping entities to constituents in the treebank. Crucial to this approach is a modification of the Penn Treebank guidelines and the characterization of entities as relation components, which allows the integration of the entity annotation with the syntactic structure while retaining the capacity to annotate and extract more complex events.

1 Introduction

A great deal of annotation effort for many different corpora has been devoted to annotation for entities and syntactic structure (treebanks). However, previous efforts at treebanking have largely been independent of the constituency of entities, and previous efforts at entity annotation have likewise been independent of corresponding layers of syntactic structure. We describe here a corpus being developed for biomedical information extraction with levels of both entity annotation and treebank annotation, with a goal that entities can be mapped to constituents in the treebank.

We are collaborating with researchers in the Division of Oncology at The Children's Hospital of Philadelphia, for the purpose of automatically mining the corpus of cancer literature for those associations that link specified variations in individual genes with known malignancies. In particular, we are interested in extracting three entities (Gene, Variation event, and Malignancy) in the following relationship: Gene X with genomic Variation event Y is correlated with Malignancy Z. For example, *WT1 is deleted in Wilms Tumor #5*. In addition, Variation events are themselves relations, consisting of entities representing different aspects of a Variation event.

Mapping entities to treebank constituents is a desirable goal since the entities can then be viewed as semantic types associated with syntactic constituents, and we expect that automated analyses of these related levels will interact in a mutually reinforcing and beneficial way for development of statistical taggers.

In this paper we describe aspects of the entity and treebank annotation that allow this mapping to be largely successful. Potentially large entities that would otherwise cut across syntactic constituents are decomposed into components of a relation. While this is worthwhile by itself on conceptual grounds for entity definition, and was in fact not done for reasons of mapping to syntactic constituents, it makes such a mapping easier. The treebank annotation has been modified from the Penn Treebank guidelines in various ways, such as greater structure for prenominal modifiers. Again, while this would have been done regardless of the mapping of entities, it does make such a mapping more successful.

Previous work on integrating syntactic structure with entity information, as well as relation infor-

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 21–28,
Ann Arbor, June 2005.

mation, is described in (Miller et al., 2000). Our work is in much the same spirit, although we do not integrate relation annotation into the syntactic trees. PubMed abstracts are quite different from the newswire sources used in that earlier work, with several consequences discussed throughout, such as the use of discontinuous entities.

Section 2 discusses some of the main issues around the development of the guidelines for entity annotation, and Section 3 discusses some of the changes that have been made for the treebank guidelines. Section 4 describes the annotation workflow and the resulting merged representation. Section 5 evaluates the mapping between entities and constituents, and Section 6 is the conclusion.

2 Guidelines for Entity Annotation

Here we give a summary of the main features of our annotation guidelines. We have been influenced in this by the annotation guidelines for the Automatic Content Extraction (ACE) project (Consortium, 2004).[1] However, our source materials are medical abstracts from PubMed[2], and important differences between the domains have required significant changes and additions to many definitions, guidelines, and procedures.

Most obviously, the vocabulary is very different. Many of the tokens in our source texts are chemical terms with a complex productive morphology, and a certain number are unique in PubMed. Many others are strings of notation, like *S37F*, often containing relevant entity references that must be isolated (*S*, *37*, and *F*). And even apart from these, we are looking at a very different dialect of English from that used by the Wall Street Journal and the Associated Press. Annotation of English newswire requires native English competency; entity annotation of biomedical English requires a background in biology as well.

The entity instances in the text are also qualitatively different. Instead of individual pieces of the physical or social universe – *Emanuel Sosa, the Eiffel Tower, the man in the yellow hat* – we have abstractions, categories that are not to be confused with their instantiations: *neuroblastoma*, *K-ras* (a gene), *codon 42*.[3] We are not currently annotating pronominal or other forms of coreference.

2.1 Entities Annotated

2.1.1 Gene Entity

For the sake of this project the definition for "Gene Entity" has two significant characteristics. First, as just mentioned, "Gene" refers to a conceptual entity as opposed to the specific manifestation of a gene (e.g., not the "K-ras" in some specific cell in some individual, but an abstraction that cannot be pointed to).

Second, "Gene" refers to a composite entity as opposed to the strict biological definition. There are often ambiguities in the usage of the entity names. I is sometimes unclear as to whether the gene or protein is being referenced, and the same name can refer to the gene or the protein at different locations in the same document. In a similar way as the ACE project allows "geopolitical" entities to have different roles, such as "location" or "organization", we consider a "Gene" to be a composite entity that can have different roles throughout a document. Therefore, Gene entity mentions can have types Gene-generic, Gene-protein, and Gene-RNA.

2.1.2 Variation Events as Relations

As mentioned in the introduction, Variation events are relations between entities representing different aspects of a Variation; specifically, a Variation is a relationship between two or more of the following entities: Type (e.g., *point mutation*, *translocation*, or *inversion*), Location (e.g., *codon 14*, *1p36.1*, or *base pair 278*), Original-State and Altered-State (e.g., *Thymine*).

The entities as such are independent and unconnected. We add a level of *relation* to annotate the associations between them: For example, the text fragment *a single nucleotide substitution at codon 249, predicting a serine to cysteine amino acid substitution (S249C)* contains the entities:

Variation-type substitution

Variation-location codon 249

[1] Another source of influence is previous work in annotation for biomedical information extraction, such as (Ohta et al., 2002). Space prevents adequate discussion of here of the differences.

[2] http://www.ncbi.nlm.nih.gov/entrez/

[3] This domain shows no such clear distinction between Name and Nominal mentions as in the texts covered by ACE.

Variation-state-original serine

Variation-state-altered cysteine

These entities are annotated individually but are also collected into a single Variation relation.

It is also possible for a Variation relation to arise from a more compact collection of entities. For example, the text *S249C* consists of three entities collected into a Variation relation:

Variation-location 249

Variation-state-original S

Variation-state-altered C

These four components represent the key elements necessary to describe any genomic variation event. Variations are often underspecified in the literature. For example, the first relation above has all four components while the second is missing the Variation-type. Characterizing individual Variations as relations among such components provides us with a great deal of flexibility.

The "Gene" entities are analogous to the ACE geopolitical entity, in that the second part of the entity names ("-RNA", "-generic","-protein") disambiguates the metonymy of the "Gene". The subtypes of the Variation entities, in contrast, indicate different kinds of entities in their own right, which can also function as components of a Variation relation.

2.1.3 Malignancy

The Malignancy annotation guidelines were under development during the annotation of the corpus described here. While they have since been more completely defined, they are not included as part of the annotated files discussed here, and so are not further discussed in this paper.

2.2 Discontinuous Entities

We have introduced a mechanism we call "chaining" to annotate discontinuous entities, which may be more common in abstracts than in full text because of the pressure to reduce word count. For example, in *K- and N-ras* there are two entities, *K-ras* and *N-ras*, of which only the second is a solid block of text. Our entity annotators are allowed to change the tokenization if necessary to isolate the components of *K-ras*:

text `K- and N-ras`

original tokenization `[K-][and][N-ras]`

		Multiple Tokens	
Entity Type	Single Tokens	Non-chains	Chains
Gene-generic	104	6	0
Gene-protein	921	349	6
Gene-RNA	1987	156	36
Var-location	95	445	125
Var-state-orig	151	5	0
Var-state-altered	162	10	0
Var-type	235	271	1

Table 1: Entity Instances

modified tokenization

`[K][-][and][N][-][ras]`

entity annotation

1. `K- ... ras` (chain with separated tokens)
2. `N-ras` (contiguous tokens)

2.3 Entity Frequencies

Table 1 shows the number of instances of each of the entity types in the 318 abstracts, discussed further in Section 4, that have been both entity annotated and treebanked. We separate the entities into single-token and multiple-token categories since it is only the multiple-token categories that raise an issue for mapping constituents.

3 Treebank Annotation

The Penn Treebank II guidelines (Bies et al., 1995) were followed as closely as possible, but the nature of the biomedical corpus has made some changes necessary or desirable. We have also taken this opportunity to address several long-standing issues with the original set of guidelines, with regard to NP structure in particular. This has resulted in the introduction of one new node label for sub-NP nominal substrings (NML). One additional empty category (*P*) has been introduced in order to improve the match-up of chained entity categories with treebank nodes. It is used as a placeholder to represent distributed modification in nominals and does not represent the trace of movement.

3.1 Tokenization/Part-of-Speech

We have also adopted several changes in word-level tokenization, leading to a number of part-of-speech and structural differences as well. Many hyphenated words are now treated as separate tokens (*New York - based* would be four tokens, for example). These hyphens now have the part-of-speech tag HYPH. If the separated prefix is a morphological unit that does not exist as a free-standing word, it has the part-of-speech tag AFX. With chemical names and scientific notation in the biomedical corpus in particular, spaces and punctuation may occur within a single "token", which will have a single POS tag.

3.2 Right-Branching Default

We assume a default binary right-branching structure under any NP and NML node. Each daughter of the phrase (whether a single token or itself a constituent node) is assumed to have scope over everything to its right. This means that every daughter also forms a constituent with everything to its right. This assumption makes the annotation process for multi-token nominals less complex and the resulting trees more legible, but still allows us to readily derive constituent nodes not explicitly represented. For example, in

```
(NP (JJ primary) (NN liver)
    (NN cancer))
```

we assume that "liver cancer" is a constituent, and that "primary" has scope over it.

So, although we do not show the intermediate nodes explicitly in our annotation, our assumed structure for this NP could be derived as

```
(NP (JJ primary)
    (newnode (NN liver)
          (newnode (NN cancer))))
```

As discussed in Section 5, entities sometimes map to such implicit constituents, and a node needs to be added to make the constituent explicit so the the entity can be mapped to it.

3.3 New Node Level for Non-Right-Branching: NML

We use the NML node label to mark nominal subconstituents that do not follow the default binary right-branching structure. Any two or more non-final elements that form a constituent are bound together by NML.

```
(NP (NML (NN human)
         (NN liver)
         (NN tumor))
    (NN analysis))
```

3.4 New Empty Category for Distributed Readings within NP: *P*

As discussed in Section 2.2, discontinuous entities are annotated using the "chaining" mechanism. Analogously, we have introduced a placeholder, *P*, for distributed material in the treebank. It is used exclusively in coordinated nominal structures, placed in coordinated elements that are missing either a distributed head or a distributed premodifier. In *K- and N-ras*, the coordinated premodifier *K-* is missing the distributed head *ras*, so the placeholder *P* is inserted after *K-* and coindexed with *ras*:

```
(NP (NP (NN K) (HYPH -)
        (NML-1 (-NONE- *P*)))
    (CC and)
    (NP (NN N) (HYPH - )
        (NML-1 (NN ras))))
```

This creates constituent nodes *K-ras* and *N-ras* that align with the entities being represented by chaining.[4]

4 Annotation Process

The annotation process comprises the following steps: Paragraph and sentence annotation (including the delimitation of irrelevant text such as author names); tokenization; entity annotation; part-of-speech (POS) annotation; treebanking; merged representation.

Entity annotation precedes POS annotation, since the entity annotators often have to correct the tokenization, which affects the POS labels. For example, *nephro- and hepatocarcinoma* refers to two entities, *nephrocarcinoma* and *hepatocarcinoma*, and so the entity annotator would split *hepatocarcinoma* into two tokens, for chaining *nephro* and *carcinoma*

[4] In spite of the apparent similarity between *P* and right node raising structures (*RNR*), they are not interchangeable as the shared element often occurs to the left rather than the right (e.g., *codon 12 or 13* in Section 5.3).

(see Section 2.2). Since the entity annotators are not qualified for POS annotation, doing POS annotation after entity annotation allows the POS annotators to annotate any such tokenization changes.

Treebank annotation uses the same tokenization as for the corresponding entity file. Continuing the above example, the treebank file would have separate tokens for *hepato* and *carcinoma*. Note that this would be the case even if we did not have the goal of mapping entities to constituents. It arises from the more minimal requirement of maintaining identical tokenization in the treebank and entity files, and so leads to changes in treebank annotation such as discussed in Section 3.4.

All of the annotation steps except entity annotation use automated taggers (or a parser in the case of treebanking),[5] producing annotation that then gets hand-corrected.

The use of the parser for producing a parse for correction by the treebankers include a somewhat unusual feature that arises from our parallel entity and treebank annotation. The parser that we are using, (Bikel, 2004),[6] allows prebracketing of parts of the parser input, so that the parser will respect the prebracketing. We use this ability to prebracket entities, which can also help to disambiguate the constituencies for prenominal modifiers, which can often be unclear for annotators without a medical background. For example, the input to the parser might contain something like:

```
...(NN activation)
   (IN of)
   (PRP$ its)
   (* (NN tyrosine)
      (NN kinase) )
   (NN activity)...
```

indicating by the `(* )` that *tyrosine kinase* should be a constituent. (It is a Gene-protein.)

Our first release of data, PennBioIE Release 0.9 (`http://bioie.ldc.upenn.edu/publications`), contains 1157 oncology PubMed abstracts, all annotated for entities and POS, of which 318 have also been treebanked. The website also contains full documentation for the various annotation guidelines mentioned in this paper.

[5] Entity taggers have been developed (McDonald et al., 2004) but have not yet been integrated into the project.

[6] Available at `http://www.cis.upenn.edu/~dbikel/software.html#stat-parser`

```
;sentence 4 Span:331..605
;In the present study, we screened for
;the K-ras exon 2 point mutations in a
;group of 87 gynecological neoplasms
;(82 endometrial carcinomas, four
;carcinomas  of the uterine cervix and
;one uterine carcinosarcoma) using the
;non-isotopic  PCR-SSCP-direct
;sequencing techniques.
;[373..378]:gene-rna:"K-ras"
;[379..385]:variation-location:"exon 2"
;[386..401]:variation-type:
            "point mutations"
(SENT
  (S
    (PP (IN:[331..333] In)
      (NP (DT:[334..337] the)
          (JJ:[338..345] present)
          (NN:[346..351] study)))
    (,:[351..352] ,)
    (NP-SBJ (PRP:[353..355] we))
    (VP (VBD:[356..364] screened)
      (PP-CLR (IN:[365..368] for)
        (NP (DT:[369..372] the)
            (NN:[373..378] K-ras)
            (NML (NN:[379..383] exon)
                 (CD:[384..385] 2))
            (NN:[386..391] point)
            (NNS:[392..401] mutations)))
      (PP (IN:[402..404] in)
        (NP
          (NP (DT:[405..406] a)
              (NN:[408..413] group))
          (PP (IN:[414..416] of)
            (NP (CD:[417..419] 87)
                (JJ:[420..433]
                    gynecological)
                (NNS:[434..443]
                     neoplasms)
[...]
```

Figure 1: Example .mrg file

4.1 Example of Merged Output

The 318 files that have been both treebanked and entity annotated are also available in a merged "`.mrg`" format. The treebank and entity annotations are both stand-off, referring to character spans in the same source file, and we take advantage of this so that the merged representation relates the entities and constituents by these spans. Figure 1 shows a fragment of one such `.mrg` file.

This `.mrg` file excerpt shows the text of sentence 4 in the file, which spans the character offsets 331..605. Each entity is listed by span (which can in-

clude several tokens), entity type, and the text of the entity. The treebank part is the same basic format as the `.mrg` files from the Penn Treebank, except that each terminal has the format

```
(POSTag:[from..to] terminal)
```

where `[from..to]` is that terminal's span in the source file.

The first entity listed, *K-ras*, is a Gene-RNA entity with span `[373..378]`, which corresponds to the single token:

```
(NN:[373..378] K-ras)
```

The second entity, *exon 2*, is a Variation-location with span `[379..385]`, which corresponds to the two tokens:

```
(NN:[379..383] exon)
(CD:[384..385] 2)
```

The third entity, *point mutations*, is a Variation-type with span `[386..401]`, which corresponds to the two tokens:

```
(NN:[386..391] point)
(NNS:[392..401] mutations)
```

By including the terminal span information in the treebank, we make explicit how the tokens that make up the entities are treated in the treebank representation.

5 Entity-Constituent Mapping

One of our goals for the release of the corpus is to allow users to choose how they wish to handle the integration of the entity and treebank information. By providing the corresponding spans for both aspects of the annotation, we provide the raw material for any integrated approach.

We therefore do not attempt to force the entities and constituents to line up perfectly. However, given the parallel annotation just illustrated, we can analyze how close we come to the ideal of the entities behaving as semantic types on syntactic constituents.

5.1 Mapping Categories

Leaving aside chains for the moment, we categorize each entity/treebank mapping in one of three ways:

Exact match There is a node in the tree that yields exactly the entity. For example, the entity *exon 2* in Figure 1

```
;[379..385]:variation-location:
            "exon 2"
```

corresponds exactly to the `NML` node in Figure 1

```
(NML (NN:[379..383] exon)
     (CD:[384..385] 2))
```

Missing node There is no node in the tree that yields exactly that entity, but it is possible to add a node to the tree that would yield the entity. A common reason for this is that the default right branching treebank annotation (Section 3.2) does not make explicit the required node.

For example, the entity *point mutations* in Figure 1

```
;[386..401]:variation-type:
            "point mutations"
```

does not correspond to a node in the relevant part of the tree:

```
(NP (DT:[369..372] the)
    (NN:[373..378] K-ras)
    (NML (NN:[379..383] exon)
         (CD:[384..385] 2))
    (NN:[386..391] point)
    (NNS:[392..401] mutations))
```

However, it is possible to insert a node into the tree to yield exactly the entity:

```
(NP (DT:[369..372] the)
    (NN:[373..378] K-ras)
    (NML (NN:[379..383] exon)
         (CD:[384..385] 2))
    (newnode (NN:[386..391] point)
             (NNS:[392..401]
                    mutations)))
```

Note that this node corresponds exactly to the implicit constituency assumed by the right branching rule. For our own internal research purposes we have generated a version of the treebank with such nodes added, although they are not in the current release.

Crossing The most troublesome case, in which the entity does not match a node in the tree and also cuts across constituent boundaries, so it is not even possible to add a node yielding the entity. Typically this

Entity Type	Total	Exact Match	Miss-ing	Cross-ing
Gene-generic	6	4	1	1
Gene-protein	349	236	103	10
Gene-RNA	156	115	35	6
Var-location	445	348	68	29
Var-state-orig	5	3	1	1
Var-state-altered	10	8	0	2
Var-type	271	123	142	6
Total	1242	837	350	55

Table 2: Matching Status of Non-Chained Multiple Token Instances

Entity Type	Total	Exact Match	Not Exact Match
Gene-generic	0	0	0
Gene-protein	6	4	2
Gene-RNA	36	29	7
Var-location	125	103	22
Var-state-orig	0	0	0
Var-state-altered	0	0	0
Var-type	1	0	1
Total	168	136	32

Table 3: Matching Status of Chained Multiple Token Instances

is due to an entity containing text corresponding to a prepositional phrase. For example, the sentence

```
One ER showed a G-to-T mutation in
the second position of codon 12
```

has the entity

```
[1280..1307]:variation-location:
            "second position
             of codon 12"
```

The relevant part of the corresponding tree is

```
(PP-LOC (IN:[1272..1274] in)
  (NP
    (NP (DT:[1276..1279] the)
        (JJ:[1280..1286] second)
        (NN:[1287..1295] position))
    (PP (IN:[1296..1298] of)
        (NP (NN:[1299..1304] codon)
            (CD:[1305..1307] 12)))))
```

Due to the inclusion of the determiner in the NP *the second position*, while it is absent from the entity definition which does include the following PP, it is not possible to add a node to the tree yielding exactly *second position of codon 12*.[7] It is possible to relax the requirements on exact match to include the determiner.[8]

However, one of our initial goals in this investigation was to determine whether this sort of limited crossing is indeed a major source of the mapping mismatches.

5.2 Overall Mapping Results

Table 2 is a breakdown of how well the (non-chain) entities can be mapped to constituents. Here we are concerned only with entities that consist of multiple tokens, since single-token entities can of course map directly to the relevant token.

The number of crossing cases is relatively small. One reason for this is the use of relations for breaking potentially large entities into component parts, since the component entities either already map to an entity or can easily be made to do so by making implicit constituents explicit to disambiguate the tree structure. The crossing cases tend to be ones in which the entities are in a sense a bit too "big", such as including a prepositional phrase.[9]

[7] The inclusion of the PP in an entity can be a problem for the constituent mapping even aside from the determiner issue. It is possible for the PP, such as *of codon 12*, to be followed by another PP, such as *in K-ras*. Since all PPs are attached at the same level, *of codon 12* and *in K-ras* are sisters, and so, even if the determiner was included in the entity name, there is no constituent consisting of just *the second position of codon 12*. However, in that case it is then possible to add a node yielding the NP and first PP. A similar issue sometimes arises when attempting to relate Propbank arguments to tree constituents.

[8] Another alternative would be to modify the treatment of noun phrases and determiners in the treebank annotation to be more akin to DPs. However, this has proved to be an impractical addition to the annotation process.

[9] As discussed in Section 4, we are prebracketing entities in the parses prepared for the treebankers to correct. There are two possibilities for how the entities can therefore ever cross treebank constituents: (1) the treebank annotation was done before we started doing such prebracketing, so the treebank annotator was not aware of the entities, or (2) the prebracketing was in-

5.3 Chained Entities

Table 3 shows the matching status of multiple token instances that are also chains (and so were not included in Table 2). The presence of chains is mostly localized to certain entity types, and the mapping is mostly successful. Variation-location contains many of the chains due to the occurrences of phrases such as *codon 12 or 13*, which map exactly to the corresponding use of the `*P*` placeholder, such as:

```
(NP (NP
      (NML-1 (NN codon))
      (CD 12))
    (CC or)
    (NP
      (NML-1 (-NONE- *P*))
      (CD 13)))
```

Cases that do not map exactly are ones in which the syntactic context does not permit the use of the placeholder *P*. For example, the text *specific codons (12, 13, and 61)*, has three discontinuous entities (*codons..12*, *codons..13*, *codons..61*), but the parenthetical context does not permit using the placeholder `*P*`:

```
(NP (JJ specific) (NNS codons)
  (PRN (-LRB- -LRB-)
    (NP (NP (CD 12))
        (, ,)
        (NP (CD 13))
        (, ,) (CC and)
        (NP (CD 61)))
    (-RRB- -RRB-)))
```

and so this example contains three mismatches.

6 Conclusion

We have described here parallel syntactic and entity annotation and how changes in the guidelines facilitate a mapping between entities and syntactic constituents. Our main purpose in this paper has been to investigate the success of this mapping. As Tables 2 and 3 show, once we make explicit the implicit right-branching binary structure, only 6.2%[10] of the entities cannot be mapped directly to a node in the tree. It also appears likely that a significant percentage of even the non-matching cases can match as well, with a slight relaxation of the matching requirement (e.g., allowing entities to have an optional determiner).

We view this in part as a successful experiment illustrating how both linguistic content and entity annotation can be enhanced by their interaction. We expect this enhancement to be useful both for biomedical information extraction in particular and more generally for the development of statistical systems that can take into account different levels of annotation in a mutually beneficial way.

Acknowledgements

The project described in this paper is based at the Institute for Research in Cognitive Science at the University of Pennsylvania and is supported by grant EIA-0205448 from the National Science Foundation's Information Technology Research (ITR) program. We would like to thank Yang Jin, Mark Liberman, Eric Pancoast, Colin Warner, Peter White, and Scott Winters for their comments and assistance, as well as the invaluable feedback of all the annotators listed at `http://bioie.ldc.upenn.edu/index.jsp?page=aboutus.html`.

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II Style, Penn Treebank Project. Tech report MS-CIS-95-06, University of Pennsylvania, Philadelphia, PA.

Daniel M. Bikel. 2004. *On the Parameter Space of Lexicalized Statistical Parsing Models*. Ph.D. thesis, Department of Computer and Information Sciences, University of Pennsylvania.

Linguistic Data Consortium. 2004. Annotation guidelines for entity detection and tracking (edt), version 4.2.6 200400401. http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF.

Ryan McDonald, Scott Winters, Mark Mandel, Yang Jin, Pete White, and Fernando Pereira. 2004. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 22(20):3249–3251.

Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *6th Applied Natural Language Processing Conference*.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, and Jun'ici Tsuji. 2002. The GENIA corpus: An annotated corpus in molecular biology domain. In *Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology*.

deed done, but the treebank annotator could not abide by the resulting tree and modified the parser output accordingly.

[10] 1410 total multiple token entities, both chained and non-chained, with 87 cases that cannot be mapped (55 crossing, 32 chained non-exact match).

Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives

Nikhil Dinesh and **Alan Lee** and **Eleni Miltsakaki** and **Rashmi Prasad** and **Aravind Joshi**
University of Pennsylvania
Philadelphia, PA 19104 USA
{nikhild,aleewk,elenimi,rjprasad,joshi}@linc.cis.upenn.edu

Bonnie Webber
University of Edinburgh
Edinburgh, EH8 9LW Scotland
bonnie@inf.ed.ac.uk

Abstract

The annotations of the Penn Discourse Treebank (PDTB) include (1) discourse connectives and their arguments, and (2) *attribution* of each argument of each connective and of the relation it denotes. Because the PDTB covers the same text as the Penn TreeBank WSJ corpus, syntactic and discourse annotation can be compared. This has revealed significant differences between syntactic structure and discourse structure, in terms of the arguments of connectives, due in large part to attribution. We describe these differences, an algorithm for detecting them, and finally some experimental results. These results have implications for automating discourse annotation based on syntactic annotation.

1 Introduction

The overall goal of the Penn Discourse Treebank (PDTB) is to annotate the million word WSJ corpus in the Penn TreeBank (Marcus et al., 1993) with a layer of discourse annotations. A preliminary report on this project was presented at the 2004 workshop on *Frontiers in Corpus Annotation* (Miltsakaki et al., 2004a), where we described our annotation of discourse connectives (both explicit and implicit) along with their (clausal) arguments.

Further work done since then includes the annotation of *attribution*: that is, who has expressed each argument to a discourse connective (the writer or some other speaker or author) and who has expressed the discourse relation itself. These ascriptions need not be the same. Of particular interest is the fact that attribution may or may not play a role in the relation established by a connective. This may lead to *a lack of congruence between arguments at the syntactic and the discourse levels*. The issue of congruence is of interest both from the perspective of annotation (where it means that, even within a single sentence, one cannot merely transfer the annotation of syntactic arguments of a subordinate or coordinate conjunction to its discourse arguments), and from the perspective of inferences that these annotations will support in future applications of the PDTB.

The paper is organized as follows. We give a brief overview of the annotation of connectives and their arguments in the PDTB in Section 2. In Section 3, we describe the annotation of the attribution of the arguments of a connective and the relation it conveys. In Sections 4 and 5, we describe mismatches that arise between the discourse arguments of a connective and the syntactic annotation as provided by the Penn TreeBank (PTB), in the cases where all the arguments of the connective are in the same sentence. In Section 6, we will discuss some implications of these issues for the theory and practice of discourse annotation and their relevance even at the level of sentence-bound annotation.

2 Overview of the PDTB

The PDTB builds on the DLTAG approach to discourse structure (Webber and Joshi, 1998; Webber et al., 1999; Webber et al., 2003) in which connectives are discourse-level predicates which project predicate-argument structure on a par with verbs at

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 29–36,
Ann Arbor, June 2005.

the sentence level. Initial work on the PDTB has been described in Miltsakaki et al. (2004a), Miltsakaki et al. (2004b), Prasad et al. (2004).

The key contribution of the PDTB design framework is its *bottom-up approach* to discourse structure: Instead of appealing to an abstract (and arbitrary) set of discourse relations whose identification may confound multiple sources of discourse meaning, we start with the annotation of discourse connectives and their arguments, thus exposing a clearly defined level of discourse representation.

The PDTB annotates as *explicit discourse connectives* all subordinating conjunctions, coordinating conjunctions and discourse adverbials. These predicates establish relations between two *abstract objects* such as events, states and propositions (Asher, 1993).[1]

We use Conn to denote the connective, and Arg1 and Arg2 to denote the textual spans from which the abstract object arguments are computed.[2] In (1), the subordinating conjunction *since* establishes a temporal relation between the event of the earthquake hitting and a state where no music is played by a certain woman. In all the examples in this paper, as in (1), Arg1 is italicized, Arg2 is in boldface, and Conn is underlined.

(1) *She hasn't played any music* since **the earthquake hit**.

What counts as a legal argument? Since we take discourse relations to hold between *abstract objects*, we require that an argument contains at least one clause-level predication (usually a verb – tensed or untensed), though it may span as much as a sequence of clauses or sentences. The two exceptions are nominal phrases that express an event or a state, and discourse deictics that denote an abstract object.

[1]For example, discourse adverbials like *as a result* are distinguished from clausal adverbials like *strangely* which require only a single abstract object (Forbes, 2003).

[2]Each connective has exactly two arguments. The argument that appears in the clause syntactically associated with the connective, we call Arg2. The other argument is called Arg1. Both Arg1 and Arg2 can be in the same sentence, as is the case for subordinating conjunctions (e.g., *because*). The linear order of the arguments will be Arg2 Arg1 if the subordinate clause appears sentence initially; Arg1 Arg2 if the subordinate clause appears sentence finally; and undefined if it appears sentence medially. For an adverbial connective like *however*, Arg1 is in the prior discourse. Hence, the linear order of its arguments will be Arg1 Arg2.

Because our annotation is on the same corpus as the PTB, annotators may select as arguments textual spans that omit content that can be recovered from syntax. In (2), for example, the relative clause is selected as Arg1 of *even though*, and its subject can be recovered from its syntactic analysis in the PTB. In (3), the subject of the infinitival clause in Arg1 is similarly available.

(2) Workers described "clouds of blue dust" *that hung over parts of the factory* even though **exhaust fans ventilated the air**.

(3) The average maturity for funds open only to institutions, considered by some *to be a stronger indicator* because **those managers watch the market closely**, reached a high point for the year – 33 days.

The PDTB also annotates *implicit connectives* between adjacent sentences where no explicit connective occurs. For example, in (4), the two sentences are contrasted in a way similar to having an explicit connective like *but* occurring between them. Annotators are asked to provide, when possible, an explicit connective that best describes the relation, and in this case *in contrast* was chosen.

(4) *The $6 billion that some 40 companies are looking to raise in the year ending March 21 compares with only $2.7 billion raise on the capital market in the previous year*. IMPLICIT - in contrast **In fiscal 1984, before Mr. Gandhi came into power, only $810 million was raised**.

When complete, the PDTB will contain approximately 35K annotations: 15K annotations of the 100 explicit connectives identified in the corpus and 20K annotations of implicit connectives.[3]

3 Annotation of attribution

Wiebe and her colleagues have pointed out the importance of ascribing beliefs and assertions expressed in text to the agent(s) holding or making them (Riloff and Wiebe, 2003; Wiebe et al., 2004; Wiebe et al., 2005). They have also gone a considerable way towards specifying how such subjective material should be annotated (Wiebe, 2002). Since we take discourse connectives to convey semantic predicate-argument relations between abstract objects, one can distinguish a variety of cases depending on the *attribution* of the discourse relation or its

[3]The annotation guidelines for the PDTB are available at `http://www.cis.upenn.edu/~pdtb`.

arguments; that is, whether the relation or arguments are ascribed to the author of the text or someone other than the author.

Case 1: The relation and both arguments are attributed to the same source. In (5), the concessive relation between Arg1 and Arg2, anchored on the connective *even though* is attributed to the speaker *Dick Mayer*, because he is quoted as having said it. Even where a connective and its arguments are not included in a single quotation, the attribution can still be marked explicitly as shown in (6), where only Arg2 is quoted directly but both Arg1 and Arg2 can be attibuted to *Mr. Prideaux*. Attribution to some speaker can also be marked in reported speech as shown in the annotation of *so that* in (7).

(5) "Now, Philip Morris Kraft General Foods' parent company is committed to the coffee business and to increased advertising for Maxwell House," says Dick Mayer, president of the General Foods USA division. "Even though **brand loyalty is rather strong for coffee**, *we need advertising to maintain and strengthen it*."

(6) *B.A.T isn't predicting a postponement* because **the units "are quality businesses and we are encouraged by the breadth of inquiries**," said Mr. Prideaux.

(7) Like other large Valley companies, Intel also noted that *it has factories in several parts of the nation*, so that **a breakdown at one location shouldn't leave customers in a total pinch**.

Wherever there is a clear indication that a relation is attributed to someone other than the author of the text, we annotate the relation with the feature value **SA** for "speaker attribution" which is the case for (5), (6), and (7). The arguments in these examples are given the feature value **IN** to indicate that they "inherit" the attribution of the relation. If the relation and its arguments are attributed to the writer, they are given the feature values **WA** and **IN** respectively.

Relations are attributed to the writer of the text by default. Such cases include many instances of relations whose attribution is ambiguous between the writer or some other speaker. In (8), for example, we cannot tell if the relation anchored on *although* is attributed to the *spokeswoman* or the author of the text. As a default, we always take it to be attributed to the writer.

Case 2: One or both arguments have a different attribution value from the relation. While the default value for the attribution of an argument is the attribution of its relation, it can differ as in (8). Here, as indicated above, the relation is attributed to the writer (annotated **WA**) by default, but Arg2 is attributed to Delmed (annotated **SA**, for some speaker other than the writer, and other than the one establishing the relation).

(8) *The current distribution arrangement ends in March 1990* , although Delmed said **it will continue to provide some supplies of the peritoneal dialysis products to National Medical**, the spokeswoman said.

Annotating the corpus with attribution is necessary because in many cases the text containing the source of attribution is located in a different sentence. Such is the case for (5) where the relation conveyed by *even though*, and its arguments are attributed to *Dick Mayer*.

We are also adding attribution values to the annotation of the implicit connectives. Implicit connectives express relations that are *inferred* by the reader. In such cases, the author *intends* for the reader to *infer* a discourse relation. As with explicit connectives, we have found it useful to distinguish implicit relations intended by the writer of the article from those intended by some other author or speaker. To give an example, the implicit relation in (9) is attributed to the writer. However, in (10) both Arg1 and Arg2 have been expressed by the speaker whose speech is being quoted. In this case, the implicit relation is attributed to the speaker.

(9) *Investors in stock funds didn't panic the weekend after mid-October's 190-point market plunge*. IMPLICIT-instead **Most of those who left stock funds simply switched into money market funds**.

(10) "People say they swim, and that may mean they've been to the beach this year," Fitness and Sports. "*It's hard to know if people are responding truthfully*. IMPLICIT-because **People are too embarrassed to say they haven't done anything**."

The annotation of attribution is currently underway. The final version of the PDTB will include annotations of attribution for all the annotated connectives and their arguments.

Note that in the Rhetorical Structure Theory (RST) annotation scheme (Carlson et al., 2003), attribution is treated as a discourse relation. We, on the other hand, do not treat attribution as a discourse

relation. In PDTB, discourse relations (associated with an explicit or implicit connective) hold between two abstracts objects, such as events, states, etc. Attribution relates a proposition to an entity, not to another proposition, event, etc. This is an important difference between the two frameworks. One consequence of this difference is briefly discussed in Footnote 4 in the next section.

4 Arguments of Subordinating Conjunctions in the PTB

A natural question that arises with the annotation of arguments of subordinating conjunctions (SUBCONJS) in the PDTB is *to what extent they can be detected directly from the syntactic annotation in the PTB*. In the simplest case, Arg2 of a SUBCONJ is its complement in the syntactic representation. This is indeed the case for (11), where *since* is analyzed as a preposition in the PTB taking an S complement which is Arg2 in the PDTB, as shown in Figure 1.

(11) Since **the budget measures cash flow**, *a new $1 direct loan is treated as a $1 expenditure*.

Furthermore, in (11), *since* together with its complement (Arg2) is analyzed as an SBAR which modifies the clause *a new $1 direct loan is treated as a $1 expenditure*, and this clause is Arg1 in the PDTB.

Can the arguments always be detected in this way? In this section, we present statistics showing that this is not the case and an analysis that shows that this lack of congruence between the PDTB and the PTB is not just a matter of annotator disagreement.

Consider example (12), where the PTB requires annotators to include the verb of attribution *said* and its subject *Delmed* in the complement of *although*. But *although* as a discourse connective denies the expectation that the supply of dialysis products will be discontinued when the distribution arrangement ends. It does **not** convey the expectation that Delmed will not say such things. On the other hand, in (13), the contrast established by *while* is between the opinions of two entities i.e., *advocates* and *their opponents*.[4]

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

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

In Section 5, we will identify additional cases. What we will then argue is that it will be insufficient to train an algorithm for identifying discourse arguments simply on the basis of syntactically analysed text.

We now present preliminary measurements of these and other *mismatches* between the two corpora for SUBCONJS. To do this we describe a procedural algorithm which builds on the idea presented at the start of this section. The statistics are preliminary in that only the annotations of a single annotator were considered, and we have not attempted to exclude cases in which annotators disagree.

We consider only those SUBCONJS for which both arguments are located in the same sentence as the connective (which is the case for approximately 99% of the annotated instances). The syntactic configuration of such relations pattern in a way shown in Figure 1. Note that it is not necessary for any of $Conn$, $Arg1$, or $Arg2$ to have a single node in the parse tree that dominates it exactly. In Figure 1 we do obtain a single node for $Conn$, and $Arg2$ but for $Arg1$, it is the set of nodes $\{NP, VP\}$ that dominate it exactly. Connectives like *so that*, and *even if* are not dominated by a single node, and cases where the annotator has decided that a (parenthetical) clausal element is not minimally necessary to the interpretation of $Arg2$ will necessitate choosing multiple nodes that dominate $Arg2$ exactly.

Given the node(s) in the parse tree that dominate $Conn$ ($\{IN\}$ in Figure 1), the algorithm we present tries to find node(s) in the parse tree that dominate $Arg1$ and $Arg2$ exactly using the operation of **tree subtraction** (Sections 4.1, and 4.2). We then discuss its execution on (11) in Section 4.3.

[4]This distinction is hard to capture in an RST-based parsing framework (Marcu, 2000). According to the RST-based annotation scheme (Carlson et al., 2003) 'although Delmed said' and 'while opponents argued' are elementary discourse units annotated in the same way: as satellites of the relation *Attribution*. RST does not recognize that satellite segments, such as the ones given above, sometimes participate in a higher RST relation along with their nuclei and sometimes not.

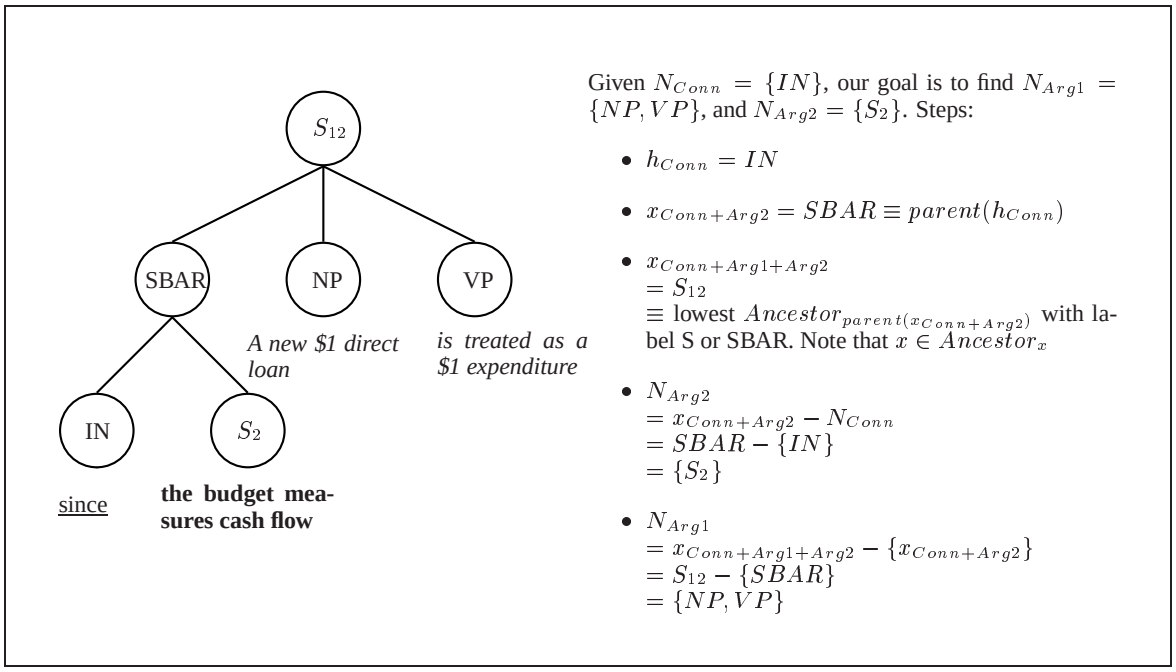


Figure 1: The syntactic configuration for (11), and the execution of the tree subtraction algorithm on this configuration.

4.1 Tree subtraction

We will now define the operation of tree subtraction the graphical intuition for which is given in Figure 2. Let T be the set of nodes in the tree.

Definition 4.1. *The ancestors of any node $t \in T$, denoted by $Ancestor_t \subseteq T$ is a set of nodes such that $t \in Ancestor_t$ and $parent(u, t) \Rightarrow ([u \in Ancestor_t] \wedge [Ancestor_u \subset Ancestor_t])$*

Definition 4.2. *Consider a node $x \in T$, and a set of nodes $Y \subset T - \{x\}$, we define the set $Z' = \{n | n \in T - \{x\} \wedge x \in Ancestor_n \wedge (\forall y \in Y, y \notin Ancestor_n \wedge n \notin Ancestor_y)\}$. Given such an x and Y, the operation of tree subtraction gives a set of nodes Z such that, $Z = \{z_1 | z_1 \in Z' \wedge (\forall z_2 \in Z', z_2 \notin (Ancestor_{z_1} - \{z_1\}))\}$*

We denote this by $x - Y = Z$.

The nodes $z \in Z$ are the highest descendants of x, which do not dominate any node $y \in Y$ and are not dominated by any node in Y.

4.2 Algorithm to detect the arguments

For any $t \in T$, let L_t denote the set of leaves(or terminals) dominated by t and for $A \subseteq T$ we denote the set of leaves dominated by A as $L_A = \bigcup_{\forall a \in A} L_a$.

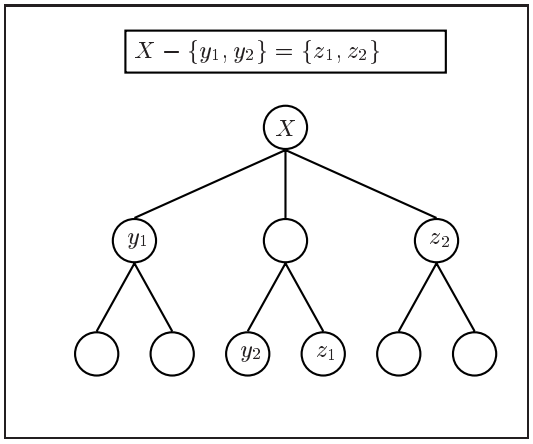


Figure 2: Tree subtraction $x - Y = Z$

For any set of leaves L we define N'_L to be a set of nodes of maximum cardinality such that $L_{N'_L} = \bigcup_{\forall n \in N'_L} L_n = L$

The set $N_L = \{n_1 | n_1 \in N'_L \wedge (\forall n_2 \in N'_L, n_2 \notin (Ancestor_{n_1} - \{n_1\}))\}$. We can think of Conn, Arg1 and Arg2 each as a set of leaves and we use N_{Conn}, N_{Arg1} and N_{Arg2} to denote the set of highest nodes which dominate them respectively.

Given N_{Conn}, our task is then to find N_{Arg1} and

N_{Arg2}. The algorithm does the following:

1. Let h_{Conn}(the head) be the last node in N_{Conn} in an in-order traversal of the tree.
2. $x_{Conn+Arg2} \equiv parent(h_{Conn})$
3. Repeat while $parent(x_{Conn+Arg2})$ has label S or SBAR, and has only two children:

   $x_{Conn+Arg2} = parent(x_{Conn+Arg2})$

   This ensures the inclusion of complementizers and subordinating conjuctions associated with the clause in Arg1. The convention adopted by the PDTB was to include such elements in the clause with which they were associated.
4. $x_{Conn+Arg1+Arg2}$ is the lowest node with label S or SBAR such that:

   $x_{Conn+Arg1+Arg2} \in Ancestor_{parent(x_{Conn+Arg2})}$
5. Repeat while $parent(x_{Conn+Arg1+Arg2})$ has label S or SBAR, and has only two children:

   $x_{Conn+Arg1+Arg2} = parent(x_{Conn+Arg1+Arg2})$
6. $N_{Arg2} = x_{Conn+Arg2} - N_{Conn}$ (tree subtraction)
7. $N_{Arg1} = x_{Conn+Arg1+Arg2} - \{x_{Conn+Arg2}\}$ (tree subtraction)

4.3 Executing the algorithm on (11)

The idea behind the algorithm is as follows. Since we may not be able to find a single node that dominates $Conn$, $Arg1$, and/or $Arg2$ exactly, we attempt to find a node that dominates $Conn$ and $Arg2$ together denoted by $x_{Conn+Arg2}$ ($SBAR$ in Figure 1), and a node that dominates $Conn$, $Arg1$ and $Arg2$ together denoted by $x_{Conn+Arg1+Arg2}$ (S_{12} in Figure 1). Note that this is an approximation, and there may be no single node that dominates $Conn$, and $Arg2$ exactly.

Given $x_{Conn+Arg2}$ the idea is to remove all the material corresponding to $Conn$ (N_{Conn}) under that node and call the rest of the material $Arg2$. This is what the operation of tree subtraction gives us, i.e., $x_{Conn+Arg2} - N_{Conn}$ which is $\{S_2\}$ in Figure 1.

Similarly, given $x_{Conn+Arg1+Arg2}$ we would like to remove the material corresponding to $Conn$ and $Arg2$ and $\{x_{Conn+Arg2}\}$ is that material. $x_{Conn+Arg1+Arg2} - \{x_{Conn+Arg2}\}$ gives us the nodes $\{NP, VP\}$ which is the desired $Arg1$.

5 Evaluation of the tree subtraction algorithm

Describing the mismatches between the syntactic and discourse levels of annotation requires a detailed analysis of the cases where the tree subtraction algorithm does not detect the same arguments as annotated by the PDTB. Hence this first set of experiments was carried out *only* on Sections 00-01 of the WSJ corpus (about 3500 sentences), which is accepted by the community to be development data.

First, the tree subtraction algorithm was run on the PTB annotations in these two sections. The arguments detected by the algorithm were classified as: (a) **Exact**, if the argument detected by the algorithm exactly matches the annotation; (b) **Extra Material**, if the argument detected contains some additional material in comparison with the annotation; and (c) **Omitted Material**, if some annotated material was not included in the argument detected. The results are summarized in Table 1.

Argument	Exact	Extra Material	Omitted Material
Arg1	82.5% (353)	12.6% (54)	4.9% (21)
Arg2	93.7% (401)	2.6% (11)	3.7% (16)

Table 1: Tree subtraction on the PTB annotations for SUBCONJS. Section 00-01(428 instances)

5.1 Analysis of the results in Table 1

5.1.1 Extra Material

There were 54 (11) cases where Arg1 (Arg2) in the PTB (obtained via tree subtraction) contained more material than the corresponding annotation in the PDTB. We describe only the cases for Arg1, since they were a superset of the cases for Arg2.

Second VP-coordinate - In these cases, Arg1 of the SUBCONJ was associated with the second of two coordinated VPs. Example (14) is the relation annotated by the PDTB, while (15) is the relation produced by tree subtraction.

(14) She became an abortionist accidentally, *and continued* because **it enabled her to buy jam, cocoa and other war-rationed goodies**.

(15) *She became an abortionist accidentally, and continued* because **it enabled her to buy jam, cocoa and other war-rationed goodies**.

Such mismatches can be either due to the fact that the algorithm looks only for nodes of type S or SBAR, or due to disagreement between the PTB and PDTB. Further investigation is needed to under-

stand this issue more precisely.[5] The percentage of such mismatches (with respect to the total number of cases of extra material) is recorded in the first column of Table 2, along with the number of instances in parentheses.

Lower Verb - These are cases of a true mismatch between the PDTB and the PTB, where the PDTB has associated Arg1 with a lower clause than the PTB. 9 of the 13 "lower verb" cases for Arg1 were due to *verbs of attribution*, as in (12). (The percentage of "lower verb" mismatches is given in the second column of Table 2, along with the number of instances in parentheses.)

Clausal Adjuncts - Finally, we considered cases where clause(s) judged not to be minimally necessary to the interpretation of Arg1 were included. (16) shows the relation annotated by the PDTB, where the subordinate clause headed by *partly because* is not part of Arg1, but the tree subtraction algorithm includes it as shown in (17).

(16) When **Ms. Evans took her job**, *several important divisions that had reported to her predecessor weren't included* partly because she didn't wish to be a full administrator.

(17) When **Ms. Evans took her job**, *several important divisions that had reported to her predecessor weren't included partly because she didn't wish to be a full administrator*.

To get an idea of the number of cases where a single irrelevant clause was included, we determined the number of instances for which pruning out one node from Arg1 resulted in an exact match. This is given in the third column of Table 2. The second row of Table 2 illustrates the same information for Arg2. Most of these are instances where irrelevant clauses were included in the argument detected from the PTB.

Argument	Second VP Coordinate	Lower Verb	One Node Pruned	Other
Arg1	16.7% (9)	24.1% (13)	31.5% (17)	27.7% (15)
Arg2	0% (0)	9.1% (1)	72.7% (8)	18.2% (2)

Table 2: Cases which result in extra material being included in the arguments.

[5]It is also possible for the PDTB to associate an argument with only the first of two coordinated VPs, but the number of such cases were insignificant.

5.1.2 Omitted Material

The main source of these errors in Arg1 are the **higher verb** cases. Here the PDTB has associated Arg1 with a higher clause than the PTB. Examples (18) and (19) show the annotated and algorithmically produced relations respectively. This is the inverse of the aforementioned *lower verb* cases, and the majority of these cases are due to the verb of attribution being a part of the relation.

(18) *Longer maturities are thought to indicate declining interest rates* because **they permit portfolio managers to retain relatively higher rates for a longer period**.

(19) Longer maturities are thought *to indicate declining* interest rates because **they permit portfolio managers to retain relatively higher rates for a longer period**.

To get an approximate idea of these errors, we checked if selecting a higher S or SBAR made the Arg1 exact or include extra material. These are the columns **Two up exact** and **Two up extra included** in Table 3. At this time, we lack a precise understanding of the remaining mismatches in Arg1, and the ones resulting in material being omitted from Arg2.

Argument	Two up exact	Two up extra included	Other
Arg1	47.6% (10)	14.3% (3)	28.1% (8)

Table 3: Cases which result in material being omitted from Arg1 as a result of excluding a higher verb

5.2 Additional experiments

We also evaluated the performance of the tree subtraction procedure on the PTB annotations on Sections 02-24 of the WSJ corpus, and the results are summarized in Table 4.

Argument	Exact	Extra Material	Omitted Material
Arg1	76.1%	17.6%	6.3%
Arg2	92.5%	3.6%	3.9%

Table 4: Tree subtraction on PTB annotations for the SUBCONJS(approx. 5K instances). Sections 02-24

Finally we evaluated the algorithm on the output of a statistical parser. The parser implementation in (Bikel, 2002) was used in this experiment and it was run in a mode which emulated the Collins (1997) parser. The parser was trained on Sections 02-21 and Sections 22-24 were used as test data, where

the parser was run and the tree subtraction algorithm was run on its output. The results are summarized in Table 5.

Argument	Exact	Extra Material	Omitted Material
Arg1	65.5%	25.2%	9.3%
Arg2	84.7%	0%	15.3%

Table 5: Tree subtraction on the output of a statistical parser (approx. 600 instances). Sections 22-24.

6 Conclusions

While it is clear that discourse annotation goes beyond syntactic annotation, one might have thought that at least for the annotation of arguments of subordinating conjunctions, these two levels of annotation would converge. However, we have shown that this is not always the case. We have also described an algorithm for discovering such divergences, which can serve as a useful baseline for future efforts to detect the arguments with greater accuracy. The statistics presented suggest that the annotation of the discourse arguments of the subordinating conjunctions needs to proceed separately from syntactic annotation – certainly when annotating other English corpora and very possibly for other languages as well.

A major source of the mismatches between syntax and discourse is the effect of attribution, either that of the arguments or of the relation denoted by the connective. We believe that the annotation of attribution in the PDTB will prove to be a useful aid to applications that need to detect the relations conveyed by discourse connectives with a high degree of reliability, as well as in constraining the inferences that may be drawn with respect to the writer's commitment to the relation or the arguments. The results in this paper also raise the more general question of whether there may be other mismatches between the syntactic and discourse annotations at the sentence level.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Press.

Daniel Bikel. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *HLT*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski, 2003. *Current Directions in Discourse and Dialogue*, chapter Building a Discourse-Tagged Corpus in the framework of Rhetorical Structure Theory, pages 85–112. Kluwer Academic Publishers.

Michael Collins. 1997. Three Generative, Lexicalized Models for Statistical Parsing. In *35th Annual Meeting of the ACL*.

Katherine Forbes. 2003. *Discourse Semantics of S-Modifying Adverbials*. Ph.D. thesis, Department of Linguistics, University of Pennsylvania.

Daniel Marcu. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, 26(3):395–448.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large scale annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004a. Annotating Discourse Connectives and their Arguments. In *the HLT/NAACL workshop on Frontiers in Corpus Annotation*, Boston, MA.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004b. The Penn Discourse Treebank. In *the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and Data Mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*, pages 105–112, Sapporo, Japan.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada, August.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalized TAG. In *ACL*, College Park, MD, June.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–87.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Janyce Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical Report TR-02-101, Department of Computer Science, University of Pittsburgh.

Investigating the Characteristics of Causal Relations in Japanese Text

Takashi Inui and **Manabu Okumura**
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259, Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
tinui@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp

Abstract

We investigated of the characteristics of in-text causal relations. We designed causal relation tags. With our designed tag set, three annotators annotated 750 Japanese newspaper articles. Then, using the annotated corpus, we investigated the causal relation instances from some viewpoints. Our quantitative study shows that what amount of causal relation instances are present, where these relation instances are present, and which types of linguistic expressions are used for expressing these relation instances in text.

1 Introduction

For many applications of natural language techniques such as question-answering systems and dialogue systems, acquiring knowledge about causal relations is one central issue. In recent researches, some automatic acquisition methods for causal knowledge have been proposed (Girju, 2003; Sato et al., 1999; Inui, 2004). They have used as knowledge resources a large amount of electric text documents: newspaper articles and Web documents.

To realize their knowledge acquisition methods accurately and efficiently, it is important to knowing the characteristics of presence of in-text causal relations. However, while the acquisition methods have been improved by some researches, the characteristics of presence of in-text causal relations are still unclear: we have no empirical study about what amount of causal relation instances exist in text and where in text causal relation instances tend to appear.

In this work, aiming to resolve the above issues, we create a corpus annotated with causal relation information which is useful for investigating what amount of causal relation instances are present and where these instances are present in text. Given some Japanese newspaper articles, we add our designed causal relation tags to the text segments. After creating the annotated corpus, we investigate the causal relation instances from three viewpoints: (i) cue phrase markers, (ii) part-of-speech information, and (iii) positions in sentences.

There are some pieces of previous work on analysis of in-text causal relations. However, although causal relation instances appear in several different ways, just a few forms have been treated in the previous studies: the verb phrase form with cue phrase markers such as in (1a) has been mainly treated. In contrast, we add our causal relation tags to several types of linguistic expressions with wide coverage to realize further analyses from above three points. Actually, we treat not only linguistic expressions with explicit cues such as in (1a), but also those without explicit cues, i.e. implicit, as in (1b), those formed by noun phrases as in (1c), and those formed between sentences as in (1d).

(1) a. 大雨-が 降っ-た ため 川-が 増水し-た。
heavy rain-NOM fall-PAST because river-NOM rise-PAST
(explicit)

b. 大雨-が 降り-、 川-が 増水し-た。
heavy rain-NOM fall-PUNC river-NOM rise-PAST
(implicit)

c. 大雨-で 川-が 増水し-た。
heavy rain-because of river-NOM rise-PAST
(noun phrase)

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 37–44,
Ann Arbor, June 2005.

d. 大雨-が 降っ-た-。 川-が 増水し-た。
heavy rain-NOM fall-PAST-PUNC river-NOM rise-PAST
(between sentences)

We apply new criteria for judging whether a linguistic expression includes a causal relation or not. Generally, it is hard to define rigorously the notion of causal relation. Therefore, in previous studies, there have been no standard common criteria for judging causal relations. Researchers have resorted to annotators' subjective judgements. Our criteria are represented in the form of linguistic templates which the annotators apply in making their judgements (see Section 3.2).

In Section 2, we will outline several previous research efforts on in-text causal relations. In Section 3 to Section 6, we will describe the details of the design of our causal relation tags and the annotation workflow. In Section 7, using the annotated corpus, we will then discuss the results for the investigation of characteristics of in-text causal relations.

2 Related work

Liu (2004) analyzed the differences of usages of some Japanese connectives marking causal relations. The results are useful for accounting for an appropriate connective for each context within the documents. However Liu conducted no quantitative studies.

Marcu (1997) investigated the frequency distribution of English connectives including "because" and "since" for implementation of rhetorical parsing. However, although Marcu's study was quantitative one, Marcu treated only explicit linguistic expressions with connectives. In the Timebank corpus (Pustejovsky et al., 2003), the causal relation information is included. However, the information is optional for implicit linguistic expressions.

Although both explicit expressions and implicit expressions are treated in the Penn Discourse Treebank (PDTB) corpus (Miltsakaki et al., 2004), no information on causal relations is contained in this corpus.

Altenberg (1984) investigated the frequency distribution of causal relation instances from some viewpoints such as document style and the syntactic form in English dialog data. Nishizawa (1997) also conducted a similar work using Japanese dialog data. Some parts of their viewpoints are overlapping with ours. However, while their studies focused on dialog data, our target is text documents. In fact, Altenberg treated also English text documents. However, our focus in this work is Japanese.

3 Annotated information

3.1 Causal relation tags

We use three tags *head*, *mod*, and *causal_rel* to represent the basic causal relation information. Our annotation scheme for events is similar to that of the PropBank (Palmer et al., 2005). An event is regarded as consisting of a head element and some modifiers. The tags *head* and *mod* are used to represent an event which forms one part of the two events held in a causal relation. The tag *causal_rel* is used to represent a causal relation between two annotated events.

Figure 1 shows an example of attaching the causal relation information to the sentence (2a), in which a causal relation is held between two events indicated (2b) and (2c). Hereafter, we denote the former (cause) part of event as e_1 and the latter (effect) part of event as e_2.

(2) a. そして、遠方からの観光客がＧＷに入って増える。
(As the Golden Week holidays come, the number of sightseers from all over begins to increase.)

b. e_1 = ＧＷに入る
(The Golden Week holidays come)

c. e_2 = 遠方からの観光客が増える
(The number of sightseers from all over begins to increase)

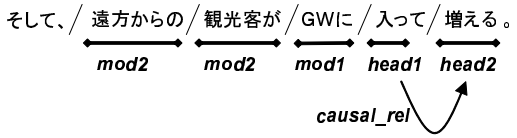


Figure 1: An example of attaching the causal relation information

The annotation process is executed as follows. First, each sentence in the text is split to some *bunsetsu*-phrase chunks[1], as shown in Figure 1 ("/" indicates a *bunsetsu*-phrase chunk boundary). Second, for each *bunsetsu*-phrase, an annotator finds the segment which represents a head element of an event,

[1]The *bunsetsu*-phrase is one of the fundamental units in Japanese, which consists of a content word (noun, verb, adjective, etc.) accompanied by some function words (particles, auxiliaries, etc.).

and he/she adds the *head* tag to the segment (see also $head_1$ and $head_2$ in Figure 1). If the event has any other elements in addition to head element, the annotator also adds the *mod* tags to the segments representing modifiers to the head element (mod_1 and mod_2 in Figure 1). The elements marked with any tags which have a common suffix number are constituents of the same event: that is, the elements marked with $head_1$ and mod_1 tags are constituents of e_1 and the elements marked with $head_2$ and mod_2 are constituents of e_2. Finally, the annotator adds the *causal_rel* tag between $head_1$ and $head_2$ as link information which indicates that the corresponding two events are held in a causal relation.

When there are any cue phrase markers helpful in recognizing causal relations such as ため (because) in (1a) , the annotator also adds the *marker* tag to their segments.

3.2 Annotation criteria

To judge whether two events represented in text are held in a causal relation or not, we apply new criteria based on *linguistic test*.

The linguistic test is a method for judging whether target linguistic expressions conforms to a given set of rules. In our cases, the target expressions are two sets of *bunsetsu*-phrase chunks. Each set represents as a whole an event which can be an argument in a causal relation, such as in (2b) and (2c) . The rules are realized as linguistic templates which are linguistic expressions including several slots.

In practice, a linguistic test is usually applied using the following steps:

1. Preparing a template.
2. Embedding the target expressions in the slots of the template to form a candidate sentence.
3. If the candidate sentence is syntactically and semantically correct, the target expressions are judged to conform to the rules. If the candidate sentence is incorrect, the targets are judged non-conforming.

In this work, we prepared eighteen linguistic templates such as in Figure 2. The square brackets indicate the slots. The symbol ⟨adv⟩ is replaced by one of three adverbs しばしば (often), 大抵 (usually), or 常に (always).

[e_1]（という）状態になれば、それに伴い、
⟨adv⟩ [e_2]（という）状態になる。

([e_2] ⟨adv⟩ happened as a result of
the fact that [e_1] happened.)

Figure 2: An example of linguistic templates

We embed two target expressions representing events in the slots of the template to form a candidate sentence. Then, if an annotator can recognize that the candidate sentence is syntactically and semantically correct, the causal relation is supposed to hold between two events. In contrast, if recognized that the candidate sentence is incorrect, this template is rejected, and the other template is tried. If all eighteen templates are rejected by the annotator, it is supposed that there is no causal relations between these two events. Note that the annotator's recognition of whether the candidate sentence is correct or incorrect, in other words, whether a causal relation is held between the two events embedded in the candidate sentence or not, is not really relevant to the author's intention.

The fundamental idea of our criteria based on linguistic test is similar to that of the criteria for annotation of implicit connectives adopted in PDTB corpus[2]. In the annotation process of the PDTB corpus, an annotator judges whether or not the explicit connective, for example, "because", relates two linguistic expressions representing events. This process is essentially the same as ours.

Three adverbs in the linguistic templates, しばしば (often), 大抵 (usually) and 常に (always), indicate a pragmatic constraint on the necessity of the relationship between any two events; the relations indicated by these words usually have a high degree of necessity. With this pragmatic constraint, we introduce an attribute to the *causal_rel* tags about the degree of necessity. For each of eighteen templates, if one judges the two target expressions as holding a causal relation by using the template with one of three adverbs, the *necessity* attribute value is added to the relation instance. If one judges the two target expressions as holding a causal relation by using the template deleting ⟨adv⟩, three adverbs, the *chance*

[2]For detail instructions of the annotation criteria in PDTB corpus, see `http://www.cis.upenn.edu/~pdtb/manual/pdtb-tutorial.pdf`.

attribute value is added.

We assume that a target expression embedded in the slot is represented by a single sentence. If an event is represented by noun phrase (NP), the following rewriting rules are applied before embedded to the slot to transform the NP into a single sentence.

- NP → NP + する
  (ex. 停電 → 停電する)
  (ex. blackout → a blackout happens)
- NP → NP + が起こる
  (ex. 地震→地震が起こる)
  (ex. earthquake → an earthquake happens)
- NP → NP + になる
  (ex. 大雨→大雨になる)
  (ex. heavy rain → it rains heavily)
- nominalized verb → verb
  (ex. 疲れ→疲れる)
  (ex. tiredness → someone gets tired)

If a head element of a target expression representing an event is conjugated, the head element is replaced by its base form before embedded to the slot.

3.3 Annotation ranges

Ideally, we should try to judge for tagging of the causal relation tags over all any event pairs in text. However, it seems that the more the distance between two events represented in text, the smaller the probability of holding a causal relation between them. Thus, we set a constraint on the ranges of judgements. If both two events are represented in the same sentence or two sentences adjacent to each other, we try judgements, if not, skip judgements. This constraint is applied only when tagging the *head* tag. A modifier and its head element are sometimes located in different sentences overtly in Japanese text when anaphora or ellipsis phenomenon occurs. In such cases, we add *mod* tags to the text segments anywhere in the text.

4 Data

We selected as text for annotation Mainichi Shimbun newspaper articles (Mainichi, 1995). In particular, we used only articles included on the social aspect domain. When adding the causal relation tags to the text, it is preferable that each annotator can understand the whole contents of the articles. The contents of social aspect domain articles seems to be familiar to everybody and are easier to understand than the contents of articles included on politics, economy domain, etc.

Furthermore, in our previous examination, it is found that as the length of articles gets longer, it is getting hard to judge which *bunsetsu*-phrase chunks represent as a whole an event. This is because as described in Section 3.3, annotators sometimes need to search several sentences for modifiers of the head element in order to add *mod* tags precisely. Therefore, we focus on social aspect domain articles which consists of less than or equal to 10 sentences. After all, we extracted 750 articles (3912 sentences) for our annotation work with above conditions.

5 Annotation workflow

Three annotators have been employed. Each annotator has added tags to the same 750 document articles independently. Two annotators of the three are linguists, and the last one is the author of this paper. We denote each annotator under anonymity, *A*, *B* and *C*. After training phase for annotators, we spent approximately one month to create a corpus annotated with causal relation information. The annotation workflow is executed efficiently using an annotation interface. Using the interface, all of annotators can add tags through only simple keyboard and mouse operations. The annotation workflow is as follows.

I. *Annotation phase:* A document article is displayed to each annotator. The sentences in the document are automatically split to *bunsetsu*-phrases by preprocessing. Some kinds of words such as connectives and verbs are highlighted to draw annotators' attention to the text segments which could represent elements in causal relation instances. The annotator finds text segments which represent causal relation instances, and then he/she adds the causal relation tags to their segments as described in Section 3.

II. *Modification phase:* After each annotator finished the annotation phase for a fixed number of document articles (in this work, 30 document articles), he/she moves to a modification phase. In this phase, first, only the segments with causal relation tags are extracted from the documents such as instances in Table 1. Then,

Table 1: Examples of tagged instances

mod_1	$head_1$	mod_2	$head_2$
6階-から (sixth floor-from)	転落する (tumble)		意識不明 (lie unconscious)
川-に (river-to)	転落 (tumble)		助け上げ (help out)
二階屋根-から (roof-from)	転落 (tumble)	頭-など-を (head-ACC)	打つ (hit)
けん銃-で (handgun-with)	撃つ (shoot)	重傷-を (heavy injury-ACC)	負う (suffer)
顔-に 火傷-を (head-DAT) (burn-ACC)	負う (suffer)		重傷 (heavy injury)
重傷-を (heavy injury-ACC)	負う (suffer)		休職する (take a sabbatical leave)

the same annotator who adds tags to the extracted segments, checks their extracted causal relation instances with attention. Since the extraction is done automatically, each annotator can check all the segments to be checked. When wrong tagged instances are found, they are corrected on the moment. After checking and correcting for all the extracted instances, the annotator moves back to the annotation phase in order to annotate a new 30 document articles set.

6 Results

6.1 Total number of tagged instances

2014 instances were tagged by the annotator *A*, 1587 instances by *B*, 1048 instances by *C*. Some examples of tagged instances are shown in Table 1.

The total numbers of tagged instances of the three annotators are quite different. Although all annotators tagged under the same annotation criteria, the annotator *A* tagged to twice as many segments as the annotator *C* did. Though this difference may be caused by some factors, we assume that the difference is mainly caused by missing judgements, since the annotators added tags to a variety of linguistic expressions, especially expressions without cue phrases.

To verify the above assumption, we again asked each annotator to judge whether or not a pair of linguistic expressions representing events is holding a causal relation. In this additional work, in order to prevent the annotators from skipping judgement itself, we present beforehand to the annotators the pairs of linguistic expressions to be judged. We presented a set of 600 pairs of linguistic expressions to each of the three annotators. All of these pairs are the causal relation instances already tagged by one or more annotators in the main work described in the previous sections.

Table 2: Inter-annotator agreement

A	*B*	*C*	$\mathcal{S}_{mixed}$	$\mathcal{S}_n$	$\mathcal{S}_c$
1	0	0	921	632	535
0	1	0	487	487	255
0	0	1	187	134	207
1	1	0	372	230	90
1	0	1	133	92	77
0	1	1	140	107	83
1	1	1	588	270	64

From the comparison between the results of the additional work and those of the main work, we found that if causal relation instances are expressed without explicit cues in text, they tend to be more frequently missed than those with explicit cues. The missing judgements on expressions without explicit cues are an important issue in the realization of more sophisticated analyses.

6.2 Inter-annotator agreement

We examined inter-annotator agreement. First, we define an agreement measure between two relation instances. Let x and y be causal relation instances tagged by two different annotators. The instance x consists of e_{1x} and e_{2x}, and y consists of e_{1y} and e_{2y}. The event e_{1x} has $head_{1x}$ as its head element. Similarly, $head_{2x}$, $head_{1y}$ and $head_{2y}$ are the head elements corresponding respectively to events e_{2x}, e_{1y} and e_{2y}. Then, we regard two instances x and y as the same instance, when $head_{1x}$ and $head_{1y}$ are located in the same *bunsetsu*-phrase and $head_{2x}$ and $head_{2y}$ are also located in the same *bunsetsu*-phrase. Using the above defined agreement measure,

we counted the number of instances tagged by the different annotators.

Table 2 shows the results. The symbol "1" in the left-hand side of Table 2 indicates that the corresponding annotator tagged to instances, and the "0" indicates not tagged. For example, the fourth row ("110") indicates that both *A* and *B* tagged to instances but *C* did not.

Let $\mathcal{S}_{mixed}$ denote a set of all tagged instances, $\mathcal{S}_n$ denote a set of all tagged instances with the *necessity* attribute value, and $\mathcal{S}_c$ denote a set of all tagged instances with the *chance* attribute value.

First, we focus on the relation instances in the set $\mathcal{S}_{mixed}$. The 1233 ($= 372 + 133 + 140 + 588$) instances are tagged by more than one annotator, and the 588 instances are tagged by all three annotators. Next, we focus on the two different contrastive sets of instances, $\mathcal{S}_n$ and $\mathcal{S}_c$. The ratio of the instances tagged by more than one annotator is small in $\mathcal{S}_c$. This becomes clear when we look at the bottom row ("111"). While the 270 instances are tagged by all three annotators in $\mathcal{S}_n$, only the 64 instances are tagged by all three annotators in $\mathcal{S}_c$.

To statistically confirm this difference, we applied the hypothesis test of the differences in population rates. The null hypothesis is that the difference of population rate is d %. As a result, the null hypothesis was rejected at 0.01 significance level when d was equal or less than 7 (p-value was equal or less than 0.00805). In general, it can be assumed that if a causal relation instance is recognized by many annotators, the instance is much reliable. Based on this assumption and the results in Table 2, reliable instances are more concentrated on the set of instances with the *necessity* attribute value than those with the *chance* attribute value.

7 Discussion

In this section, we discuss some characteristics of in-text causal relations and suggest some points for developing the knowledge acquisition methods for causal relations. Here, to guarantee the reliability of the data used for the discussion, we focus on the 699 ($= 230 + 92 + 107 + 270$) instances marked by more than one annotator with the *necessity* attribute value. We examined the following three parts: (i) cue phrase markers, (ii) the parts-of-speech of head elements, and (iii) the positions of head elements.

Table 3: The number of instances with/without cue phrase markers

with marker	219
without marker	480

Table 4: Cue phrase markers marked by annotators

marker		frequency
ため	(because)	120
で	(by)	35
結果	(result of)	5
ので	(because)	5
と	(when)	5
場合	(when)	4
ば	(if)	4
ことから	(from)	4
から	(from)	3

7.1 Cue phrase markers

While annotating the document articles with our causal relation tags, *head*, *mod*, and *causal_rel*, the annotators also marked the cue phrase markers for causal relations with the *marker* tag at the same time. We investigated a proportion of instances attached with the *marker* tag.

The result is shown in Table 3. Table 4 shows the cue phrase markers actually marked by at least one annotator [3].

It has been supposed that causal relation instances are sometimes represented with no explicit cue phrase marker. We empirically confirmed the supposition. In our case, only 30% of our 699 instances have one of cue phrase markers shown in Table 4, though this value can be dependent of the data.

This result suggests that in order to develop knowledge acquisition methods for causal relations with high coverage, we must deal with linguistic expressions with no explicit cue phrase markers as well as those with cue phrase markers.

7.2 The parts-of-speech of head elements

Next, we classified the events included in the 699 instances into two syntactic categories: the verb phrase (VP) and the noun phrase (NP). To do this, we used morphological information of their head elements. If the part-of-speech of a head is verb or adjective, the event is classified as a verb phrase. If

[3]The cue phrase markers whose frequencies are less than three are not listed due to space limitation in Table 4.

Table 5: The syntactic types

		e_1	e_2
VP	(verb) (adjective)	365	412
NP	(verbal noun) (general noun)	322	269
others		12	18

the part-of-speech of a head is noun (including general noun and verbal noun), the event is classified as a noun phrase. We used *ChaSen* [4] to get part-of-speech information.

The result is shown in Table 5. More than half events are classified as the VP. This matches our intuition. However, the number of events classified as the NP is comparable to the number of events classified as the VP; 322 events of e_1 are represented as noun phrases, and 269 events of e_2 are also represented as noun phrases.

This result is quite suggestive. To promote the current methods for knowledge acquisition to further stage, we should develop a knowledge acquisition framework applicable both to the verb phrases and to the noun phrases.

7.3 The positions of head elements

For each e_1 and e_2 included in the 699 instances, we examined the positions of their head elements in the sentences.

We consider dependency structures between *bunsetsu*-phrases in the original sentences from which causal relation instances are extracted. The dependency structures form tree structures. The *bunsetsu*-phrase located in the end of the sentence is the root node of the tree. We focus on the depth of the head element from the root node. We used *CaboCha*[5] to get dependency structure information between *bunsetsu*-phrases.

The results are shown in Figure 3 and Figure 4. Figure 3 is the result for the head elements of e_1, and Figure 4 is the result for the head elements of e_2. The letter "f" in Figure 3 and Figure 4 indicates frequency at each position. Similarly, the letter "c" indicates cumulative frequency.

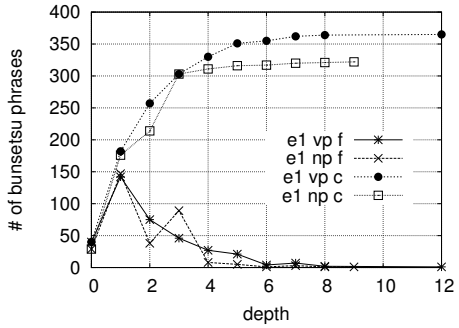


Figure 3: The positions of head elements (e_1)

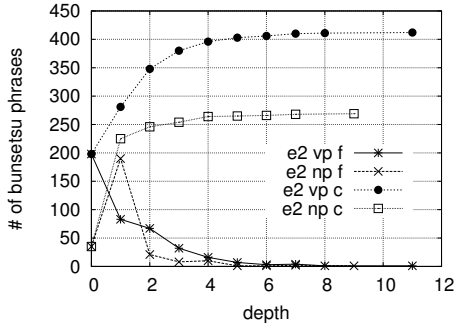


Figure 4: The positions of head elements (e_2)

In Figure 4, the 198 head elements of the events represented as a verb phrase are located in the end of the sentences, namely depth $= 0$. The 190 of the 269 events represented as a noun phrase are located in depth $= 1$. For events represented as either a verb phrase or a noun phrase, over 80% of head elements of the events are located within depth < 3. In Figure 3, similarly, over 80% of head elements of the events are located within depth < 4.

These findings suggest that the most of the events are able to be found simply by searching the *bunsetsu*-phrases located in the shallow position at the phase of causal knowledge acquisition.

7.4 Relative positions of two head elements

Finally, we examined relative positions between head elements of e_1 and e_2 where these two events are held in a causal relation. In Section 7.3, we discussed each absolute position for e_1 and e_2 by means of the notion of depth in sentences. Here, we focus on the difference (D) of the depth values between e_1 and e_2.

The result is shown in Table 6. The symbol "$e_1 \Rightarrow e_2$" in Table 6 indicates the case where the head element of e_1 is located nearer to the beginning of the

[4] Available from `http://chasen.aist-nara.ac.jp/hiki/ChaSen/`.

[5] Available from `http://chasen.org/~taku/software/cabocha/`.

Table 6: Relative positions of two head elements

		$e_1 \Rightarrow e_2$	$e_2 \Rightarrow e_1$
intra-sentential	$D = 1$	259	15
	$= 2$	152	23
	> 2	33	4
	no dep	72	
inter-sentential		141	

sentence than that of e_2. The "$e_2 \Rightarrow e_1$" indicates the opposite case. The symbol "no dep" indicates the case where neither the condition *a* nor *b* is satisfied:

a. the head element of e_2 is an ancestor of the head element of e_1.

b. the head element of e_2 is a descendant of the head element of e_1.

The symbol "inter-sentential" indicates the case where two head elements appear in different sentences.

The most instances (259 instances) are categorized into $D = 1$ on $e_1 \Rightarrow e_2$, that is, the head element of e_1 directly depends on the head element of e_2. This result matches our intuition. However, there are several other cases. For example, 152 instances are categorized into $D = 2$ on $e_1 \Rightarrow e_2$, 72 instances are categorized into "no dep". Most of the instances extracted from sentences including any parallel relations are categorized into "no dep". In this study, we consider causal relation instances as binary relation. To deal with instances categorized into "no dep" adequately, we should extend our framework to the more complex structure.

8 Conclusion

We reported our causal relation tags and the annotation workflow. Using the annotated corpus, we examined the causal relation instances in Japanese text. From our investigation, it became clear that what amount of causal relation instances are present, where these relation instances are present, and which types of linguistic expressions are used for expressing these relation instances in text.

Acknowledgement

This research is supported by the 21COE Program "Framework for Systematization and Application of Large-Scale Knowledge Resources" and the Grant-in-Aid for Creative Basic Research (13NP0301) "Language Understanding and Action Control". We would like to express our special thanks to Junji Etoh, Yoshiko Ueda, Noriko Sogoh, and Tetsuro Takahashi for helping us to create our corpus. We are grateful to the reviewers for their suggestive comments.

References

B. Altenberg. 1984. Causal linking in spoken and written English. *Studia Linguistica*, 38:1.

R. Girju. 2003. Automatic detection of causal relations for question answering. In *Proc. of the 41st ACL, Workshop on Multilingual Summarization and Question Answering*.

T. Inui. 2004. *Acquiring causal knowledge from text using connective markers*. Ph.D. thesis, Graduate School of Information Science, Nara Institute of Science and Technology.

Y. Liu. 2004. *Semantics and usages of connectives for causal relations in modern Japanese - cases of 'dakara', 'sitagatte', 'soreyue(ni)', 'sonokekka', 'sonotame(ni)'* -. Ph.D. thesis, The Graduate School of Languages and Cultures, Nagoya University.

Mainichi. 1995. Mainichi Shimbun CD-ROM version.

D. Marcu. 1997. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Department of Computer Science, University of Toronto.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. Annotating discourse connectives and their arguments. In *Proc. of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*.

S. Nishizawa and Y. Nakagawa. 1997. A method of discourse structure understanding in Japanese task-free conversation for causal conjuction. *Natural Language Processing*, 4(4):61–72. (in Japanese).

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).

J. Pustejovsky, J. M. Castaño, R. Ingria, R. Sauri, R. J. Gaizauskas, A.Setzer, G. Katz, and D. R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*, pages 28–34.

H. Sato, K. Kasahara, and K. Matsuzawa. 1999. Rertrieval [sic] of simplified causal knowledge in text and its application. In *Technical report of IEICE, Thought and Language*. (in Japanese).

A Framework for Annotating Information Structure in Discourse

Sasha Calhoun[1], **Malvina Nissim**[1], **Mark Steedman**[1] **and Jason Brenier**[2]

[1]Institute for Communicating and Collaborative Systems, University of Edinburgh, UK
Sasha.Calhoun@ed.ac.uk, {steedman,mnissim}@inf.ed.ac.uk
[2]Department of Linguistics, University of Colorado at Boulder
jbrenier@colorado.edu

Abstract

We present a framework for the integrated analysis of the textual and prosodic characteristics of information structure in the *Switchboard* corpus of conversational English. Information structure describes the availability, organisation and salience of entities in a discourse model. We present standards for the annotation of *information status* (old, mediated and new), and give guidelines for annotating *information structure*, i.e. *theme/rheme* and *background/kontrast*. We show that information structure in English can only be analysed concurrently with prosodic prominence and phrasing. This annotation, using stand-off XML in NXT, can help establish standards for the annotation of information structure in discourse.

1 Introduction

We present a framework for the integrated analysis of the textual and prosodic characteristics of information structure in a corpus of conversational English. Section 2 introduces the corpus as well as the tools we employ in the annotation process. We propose two complementary annotation efforts within this framework. The first, information status (*old, mediated, new*), expresses the *availability* of entities in discourse (Section 3). The second scheme will firstly annotate *theme/rheme*, i.e. how each intonation phrase is organised in the discourse model, and secondly *kontrast*: how *salient* the speaker wishes to make each entity, property or relation (Section 4). We will demonstrate that the perception of both of these is intimately affected by prosodic structure. In particular, the theme/rheme division affects prosodic phrasing; and information status and kontrast affect relative prosodic prominence. Therefore we also propose to annotate a subset of the corpus for this prosodic information (Section 5). In conjunction with existing annotations of the corpus, our integrated framework using NXT will be unique in the field of conversational speech in terms of size and richness of annotation.

2 Corpus and Tools

The Switchboard Corpus (Godfrey et al., 1992) consists of 2430 spontaneous phone conversations (average six minutes), between speakers of American English, for three million words. The corpus is distributed as stereo speech signals with an orthographic transcription per channel time-stamped at the word level. A third of this is syntactically parsed as part of the Penn Treebank (Marcus et al., 1993) and has dialog act annotation (Shriberg et al., 1998). We used a subset of this. In adherence with current standards, we converted all the existing annotations, and are producing the new discourse annotations in a coherent multi-layered XML-conformant schema, using NXT technology (Carletta et al., 2004).[1] This allows us to search over and integrate information from the many layers of annotation, including the

[1]Beside the NXT tools, we also used the TIGER Switchboard filter (Mengel and Lezius, 2000) for the XML-conversion. Using existing markup we automatically selected and filtered NPs to be annotated, excluding locative, directional, and adverbial NPs and disfluencies, and adding possessive pronouns. See (Nissim et al., 2004) for technical details.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 45–52,
Ann Arbor, June 2005.

sound files. NXT tools can be easily customised to accommodate different layers of annotation users want to add, including data sets that have low-level annotations time-stamped against a set of synchronized signals, multiple, crossing tree structures, and connection to external corpus resources such as gesture ontologies and lexicons (Carletta et al., 2004).

3 Information Status

Information Status describes how *available* an entity is in the discourse. We define this in terms of the speaker's assumptions about the hearer's knowledge/beliefs, and we express it by the well-known old/new distinction.[2]

3.1 Annotation Scheme

Our annotation scheme for the discourse layer mainly builds on (Prince, 1992) and (Eckert and Strube, 2001), as well as on related work on annotation of anaphoric links (Passonneau, 1996; Hirschman and Chinchor, 1997; Davies et al., 1998; Poesio, 2000). Prince defines "old" and "new" with respect to the *discourse model* as well as the *hearer*'s point of view. Considering the interaction of both these aspects, we define as *new* an entity which has not been previously referred to and is yet unknown to the hearer, and as *mediated* an entity that is newly mentioned in the dialogue but that the hearer can *infer* from the prior context.[3] This is mainly the case of generally known entities (such as "the sun", or "the Pope" (Löbner, 1985)), and *bridging* (Clark, 1975), where an entity is related to a previously introduced one. Whenever an entity is not new nor mediated is considered as *old*.

Because finer-grained distinctions (e.g. (Prince, 1981; Lambrecht, 1994)) have proved hard to distinguish reliably in practice, we organise our scheme *hierarchically*: we use the three main classes described above as top level categories for which more specific subtypes can assigned. This approach preserves a high-level, more reliable distinction while allowing a finer-grained classification that can be exploited for specific tasks.

Besides the main categories, we introduce two more classes. A category non-applicable is used for wrongly extracted markables (such as "course" in "of course"), for idiomatic occurrences, and expletive uses of "it". Traces are automatically extracted as markables, but are left unannotated. In the rare event the annotators find some fragments too difficult to understand, a category not-understood can be assigned. Entities marked as non-applicable or not-understood are excluded from any further annotation. For all other markables, the annotators must choose between old, mediated, and new. For the first two, subtypes *can* also be specified: subtype assignment is encouraged but not compulsory.

New The category new is assigned to entities that have not yet been introduced in the dialogue and that the hearer cannot infer from previously mentioned entities. No subtypes are specified for this category.

Mediated Mediated entities are inferrable from previously mentioned ones, or generally known to the hearer. We specify nine subtypes: general, bound, part, situation, event, set, poss, func_value, aggregation.[4] Generally known entities such as "the moon" or "Italy" are assigned a subtype general. Most proper nouns fall into this subclass, but the annotator could opt for a different tag, depending on the context. Also mediated are bound pronouns, such as "them" in (1), which are assigned a subtype bound.[5]

(1) [...] it's hard to raise *one child* without **them** thinking they're the pivot point of the universe.

A subtype poss is used to mark all kinds of intra-phrasal possessive relations (pre- and postnominal).

Four subtypes (part, situation, event, and set) are used to mark instances of bridging. The subtype part is used to mark part-whole relations for physical objects, both as intra- and inter-phrasal relations. (This category is to be preferred to poss whenever applicable.) The occurrence of "the door" in (2), for instance, is annotated as mediated/part.

(2) When I come *home* in the evenings my dog greets me at **the door**.

For similar relations that do not involve physical objects, i.e. if an entity is part of a situation set up by

[2] We follow Prince in using "old" rather than "given" to refer to "not-new" information, but regard the two as identical.

[3] This type corresponds to Prince's (1981; 1992) *inferrables*.

[4] Some of the subtypes are inspired by categories developed for bridging markup (Passonneau, 1996; Davies et al., 1998).

[5] All examples in this paper are from the Switchboard Corpus. The markable in question is typed in boldface; antecedents or trigger entities, where present, are in italics. For the sake of space we do not provide examples for each category (see (Nissim, 2003)).

a previously introduced entity, we use the subtype situation.[6],as for the NP "the specifications" in (3).

(3) I guess I don't really have a problem with *capital punishment*. I'm not really sure what **the exact specifications** are for Texas.

The subtype event is applied whenever an entity is related to a previously mentioned verb phrase (VP). In (4), e.g., "the bus" is triggered by *travelling around Yucatan*.

(4) We were *travelling around Yucatan*, and **the bus** was really full.

Whenever an entity referred to is a subset of, a superset of, or a member of the same set as a previously mentioned entity, the subtype set is applied.

Rarely, an entity refers to a value of a previously mentioned function, as "zero" and "ten" in (5). In such cases a subtype func-value is assigned.

(5) I had kind of gotten used to *centigrade temperature* [...] if it's between **zero** and **ten** it's cold.

Lastly, a subtype aggregation is used to classify coordinated NPs. Two old or med entities, for instance do not give rise to an old coordinated NP, unless it has been previously introduced as such. A mediated/aggregation tag is assigned instead.

Old An entity is old when it is not new nor mediated. This is usually the case if an entity is *coreferential* with an already introduced entity, if it is a generic pronoun, or if it is a personal pronoun referring to the dialogue participants. Six different subtypes are available for old entities: identity, event, general, generic, ident_generic, relative. In (6), for instance, "us" would be marked as old because it corefers with "we", and a subtype identity would also be assigned.

(6) [...] *we* camped in a tent, and uh there were two other couples with **us**.

In addition, a coreference link is marked up between anaphor and antecedent, thus creating anaphoric chains (see also (Carletta et al., 2004)). The subtype event applies whenever the antecedent is a VP. In (7), "it" is old/event, as its antecedent is the VP "educate three". As we do not extract VPs as markables, no link can be marked up.

(7) I most certainly couldn't *educate three*. I don't know how my parents did **it**.

[6]This includes elements of the thematic grid of an already introduced entity. It subsumes Passonneau's (1996) class "arg".

Also classified as old are personal pronouns referring to the dialogue participants as well as generic pronouns. In the first case, a subtype general is specified, whereas the subtype for the second is generic. An instance of old/generic is "you" in (8).

(8) up here **you** got to wait until Aug- August until the water warms up.

In a chain of generic references, the subtype ident_generic is assigned, and a coreference link is marked up. Coreference is also marked up for relative pronouns: they receive a subtype relative and are linked back to their head.

The guidelines contain a decision tree the annotators use to establish priority in case more than one class is appropriate for a given entity. For example, if a mediated/general entity is also old/identity the latter is to be preferred to the former. Similar precedence relations hold among subtypes.

To provide more robust and reliable clues in annotating bridging types (e.g. for distinguishing between poss and part), we provided replacement tests and referred to relations encoded in knowledge bases such as WordNet (Fellbaum, 1998) (for part) and FrameNet (Baker et al., 1998) (for situation).

3.2 Validation of the Scheme

Three Switchboard dialogues (for a total of 1738 markables) were marked up by two different annotators for assessing the validity of the scheme. We evaluated annotation reliability by using the Kappa statistic (Carletta, 1996). Good quality annotation of discourse phenomena normally yields a kappa (K) of about .80. We assessed the validity of the scheme on the four-way classification into the three main categories (old, mediated and new) and the non-applicable category. We also evaluated the annotation including the subtypes. All cases where at least one annotator assigned a not-understood tag were excluded from the agreement evaluation (14 markables). Also excluded were all traces (222 markables), which the annotators left unmarked. The total markables considered for evaluation over the three dialogues was therefore 1502.

The annotation of the three dialogues yielded $K = .845$ for the high-level categories, and $K = .788$ when including subtypes ($N = 1502$; $k = 2$).[7]

[7]N stands for the number of instances annotated and k for

These results show that overall the annotation is reliable and that therefore the scheme has good reproducibility. When including subtypes agreement decreases, but backing-off to the high-level categories is always possible, thus showing the virtues of a hierarchically organised scheme. Reliability tests for single categories showed that mediated and new are more difficult to apply than old, for which agreement was measured at $K = .902$, although still quite reliable ($K = .800$ and $K = .794$, respectively). Agreement for non-applicable was $K = .846$.

The annotators found the decision tree very useful when having to choose between more than one applicable subtype, and we believe it has a significant impact on the reliability of the scheme.

The scheme was then applied for the annotation of a total of 147 Switchboard dialogues. This amounts to 43358 sentences with 69004 annotated markables, 35299 of which are old, 23816 mediated and 9889 new (8127 were excluded as non-applicable, and 160 were not understood), and 16324 coreference links.

In Section 6 we use this scheme to annotate the Pie-in-the-Sky text.

3.3 Related Work

To our knowledge, (Eckert and Strube, 2001) is the only other work that explicitly refers to IS annotation. They also use a Prince's (1992)-based old/med/new distinction for annotating Switchboard dialogues. However, their IS annotation is specifically designed for salience ranking of candidate antecedents for anaphora resolution, and not described in detail. They do not report figures on inter-annotator agreement so that a proper comparison with our experiment is not feasible. Among the schemes that deal with annotation of anaphoric NPs, our scheme is especially comparable with DRAMA (Passonneau, 1996) and MATE (Davies et al., 1998). Both schemes have a hierarchical structure. In DRAMA, types of *inferrables* can be specified, within a division into conceptual (pragmatically determined) vs. linguistic (based on argument structure) inference. No annotation experiment with inter-annotator agreement figures is however reported. MATE provides subtypes for bridging relations, but they were not applied in any annotation excercise, so that reliability and distribution of categories are only based on the "core scheme" (true coreference). For a detailed comparison of our approach with related efforts on the annotation of anaphoric relations, see (Nissim et al., 2004).

the number of annotators. Unless otherwise specified, $N = 1502$ and $k = 2$ hold for all K scores reported in Section 3.

4 Information Structure

We have seen that information status describes how available an entity is in a discourse. Generally *old* entities are available, and *new* entities are not. In prosody we find that newness is highly correlated with pitch accenting, and oldness with deaccenting (Cutler et al., 1997). However, this is only one aspect of information structure. We also need to describe how speakers signal the organisation and salience of elements in discourse. Building on the work of (Vallduví & Vilkuna, 1998), as developed by (Steedman, 2000), we define two notions, *theme/rheme* structure and *background/kontrast*.

Theme/rheme structure guides how an element fits into the discourse model: if it relates back it is *thematic*; if it advances the discourse it is *rhematic*. Steedman claims that intonational phrases can mark information units (*theme* and *rheme* - though not all boundaries are realised and a unit may contain more than one phrase). The pitch contour associated with nuclear accents in themes is distinct from that in rhemes (which he identifies as L+H*LH% and H*LH% re ToBI (Beckman and Elam, 1997)), so that, where present, such boundaries disambiguate information structure. (See (9)).[8]

(9) (Q) Personally, I love hyacinths.
What kind of bulbs grow well in your area?
(A)

(In	MY	AREA)	
Bkgd	*Kont.*	*Bkgd*	*(Theme)*
(it is the	DAFFODIL)		
Bkgd	*Kont.*		*(Rheme)*

The second dimension, *kontrast*, relates to salience.[9] We expect new entities to be salient and old entities not. Therefore, if an old element is salient, or a new one especially salient, an extra meaning is implied.

[8] Annotation is as in Section 3. Words in SMALL CAPS are accented, parentheses indicate intonation phrases, including boundary tones if present. See website to hear some examples from this section.

[9] We use *kontrast* to distinguish it from the everyday use of *contrast* and the sometimes conflicting uses of *contrast* in the literature. Annotators, however, will not be given this term.

These are largely subsumed by *kontrast*, i.e. distinguishing an element from alternatives made available by the context (See (9)).

4.1 Annotation Scheme

As we have seen, in English, information structure is primarily conveyed by intonation. We therefore think it is vital for annotators to listen to the speech while annotating this structure.

4.1.1 *Theme/Rheme*

We have claimed that prosodic phrasing can divide utterances into information units. However, often theme material is entirely background, i.e., mutually known and without contrasting alternatives. Therefore, for both model theoretic and practical purposes, it is the same as background of the rheme. Accordingly, we work with a test for themehood, defining the **rheme** as any prosodic phrase that is not identifiable as a **theme**.

Annotators will mark each prosodic phrase as a *theme* if it only contains information which links the utterance to the preceding context, i.e. setting up what they're saying in relation to what's been said before. In their opinion, even if this is not the tune the speaker used, it must sound appropriate if they say it with a highly marked tune, such as L+H* LH%. For example, in (10), the phrase "where I lived" links "was a town called Newmarket" to the statement the speaker lived in England (accenting not shown). It would be appropriate to utter it with an L+H* accent on "Where" and/or "lived,", and a final LH%. So it is a theme. The same accent on "town" and/or "Newmarket" sounds inappropriate, and it advances the discussion, so it is a rheme.

(10) I lived over in England for four years
(Where I lived) *(Theme)*
(was a town called Newmarket) *(Rheme)*

4.1.2 *Background/Kontrast*

Although there is a clear link between prosodic prominence and kontrast, there are a number of disagreements about how this works which this annotation effort seeks to resolve. Some, including (Steedman, 2000), have claimed that kontrast within theme and kontrast within rheme are marked by categorically distinct pitch accents. Another view is that kontrast, also called contrastive focus or topic, only applies to *themes* that are contrastive; the head of a rheme phrase always attracts a pitch accent, it is therefore redundant to call one part kontrastive. Further, some consider kontrast within a rheme phrase only occurs when there is a clear alternative set, i.e. the distinction between broad and narrow focus, as in (9) where *daffodil* contrasts with other bulbs the speaker might grow. Again, there is controversy on whether there is an intonational difference between broad and narrow focus (Calhoun, 2004a). If these distinctions are marked prosodically, it is disputed whether this is with different pitch accents (Steedman), or by the relative height of different accents in a phrase (Rump and Collier, 1996; Calhoun, 2004b).

Rather than using the abstract notion of kontrast directly, annotators will identify discourse scenarios which commonly invoke kontrast (drawing on functions of emphatic accents from (Brenier et al., 2005)).[10] This addresses the disagreements above, while making our annotation more constrained and robust. In each case, using the full discourse context including the speech, annotators mark each content word (noun, verb, adjective, adverb and demonstrative pronoun) for the first category that applies. If none apply, they mark it as **background**.

correction The speaker's intent is to correct or clarify another just used by them or the other speaker. In (11), e.g., the speaker wishes to clarify whether her interlocutor really meant "hyacinths".

(11) (now are you sure they're **HYACINTHS**) (because that is a BULB)

contrastive The speaker intends to contrast the word with a previous one which was (a) a current topic; (b) semantically related to the contrastive word, such that they belong to a natural set. In (12), B contrasts recycling in her town "San Antonio", with A's town "Garland", from the set *places where the speakers live*.

(12) (A) I live in *Garland*, and we're just beginning to build a real big recycling center...
(B) (YEAH there's been) (NO emphasis on recycling at ALL) (in **San ANTONIO**)

[10] Emphasis can occur for two major reasons, both identified by Brenier: emphasis of a particular word or phrase, i.e. kontrast, or emphasis over a larger span of speech, conveying affective connotations such as excitement, which is not included here. (Ladd, 1996).

subset The speaker highlights one member of a more general set that has been mentioned and is a current topic. In (13), the speaker introduces "three day cares", and then gives a fact about each.

(13) (THIS woman owns *THREE day cares*) (**TWO** in Lewisville) (and **ONE** in Irving) (and she had to open **the SECOND one** up) (because her WAITING list was) (a YEAR long)

adverbial The speaker uses a focus-sensitive adverb, i.e. *only*, *even*, *always* or *especially* to highlight that word, and not another in the natural set. The adverb and/or the word can be marked. In (14), B didn't even like the "previews" of 'The Hard Way', let alone the movie.

(14) (A) I like Michael J Fox, though I thought he was crummy in 'The Hard Way'.
(B) (I didn't even like) (the **PREVIEWS**)

answer The word (or its syntactic phrase, e.g. an NP) and no other, fills to an open proposition set up in the context. It must make sense if they had only said that word or phrase. In (15), A sets up the "blooms" she can't identify, and B answers "lily".

(15) (A) We have *these blooms*, I'm not sure what they are but they come in all different colours yellow, purple, white...
(B) (I BET you) (that that's a **LILY**)

Again, in Section 6 we apply the scheme to the Pie-in-the-Sky text.

4.2 Related Work

Annotator agreement for pitch accents and prosodic boundaries, re ToBI, is about 80% and 90% respectively (Pitrelli et al., 1994). Automatic performance, using acoustic and textual features, is now above 85% accuracy (Shriberg et al., 2000). However, this does not distinguish prosodic events which occur for structural or rhythmical reasons from those which mark information structure (Ladd, 1996). (Heldner et al., 1999) try to predict focal accents. They define this minimally as the most prominent in a three-word phrase. (Hirschberg, 1993) got 80-98% accuracy using only text-based features. However, her definition of contrast was not as thorough as ours. (Hedberg and Sosa, 2001) looked at marking of ratified, unratified (old and new) and contrastive topics and foci (theme and rheme) with ToBI pitch accents. (Baumann et al., 2004) annotated a simpler information structure and prosodic events in a small German corpus.

5 Information Structure and Prosodic Structure

Much previous work, not corpus-based, draws a direct correspondence between information structure, prosodic phrasing and pitch accent type. However in real speech there are many non-semantic influences on prosody, including phrase length, speaking rate and rhythm. Information structure is rather a strong constraint on the realisation of prosodic structure (Calhoun, 2004a). Contrary to the assumption of ToBI, this structure is metrical, highly structured and linguistically relevant both within and across prosodic phrases (Ladd, 1996; Truckenbrodt, 2002).

One of our main aims is to test how such evidence can be reconciled with theories presented earlier about the relationship between information structure and prosody. Local prominence levels have been shown to aid in the disambiguation of focal adverbs, anaphoric links, and global discourse structures marked as *elaboration*, *continuation*, and *contrast* (Dogil et al., 1997). Global measures of prominence level have been linked to topic structure, corrections, and turn-taking cues (Ayers, 1994). (Brenier et al., 2005) found that *emphatic* accents realised special discourse functions such as *assessment*, *clarification*, *contrast*, *negation* and *protest* in child-directed speech. Most of these functions can be seen as conversational implicatures of kontrast, i.e. if an element is unexpectedly highlighted, this implies an added meaning. Brenier found that while pitch accents can be detected using both acoustic and textual cues; textual features are not useful in detecting emphatic pitch accents, showing there is added meaning not available from the text.

As noted in Section (4.2), inter-annotator agreement for the identification of prosodic phrase boundaries with ToBI is reasonably good. We will therefore label ToBI break indices 3 and 4 (conflated) (Beckman and Elam, 1997). Annotators will also mark the perceived level of prosodic prominence on each word using a defined scale. We are currently running a pilot experiment to identify a reasonable number of gradations of prosodic prominence, from completely unstressed and/or reduced to highly emphatic, to use for the final annotation.

[But [[[Yemen' s]$_{med/general}$ president]$_{med/poss}$]]$_{Contrastive}$ says]$_{THEME}$ [[the FBI]$_{old/identity}$ has told [him]$_{old/identity}$]$_{THEME}$ [[the explosive material]$_{med/set}$ could only have come from [[[the U.S.]$_{med/general}$, [israel]$_{med/general}$, or [[two arab countries]$_{med/set}$]$_{med/aggregation}$.]$_{Adverbial}$]$_{RHEME}$ [And to [[a former federal bomb investigator]$_{new}$,]$_{Contrastive}$]$_{THEME}$ [[that description]$_{old/event}$ suggests]$_{THEME}$ [[a powerful military-style plastic explosive C-4]$_{med/set}$]$_{Answer}$ [[that]$_{old/relative}$ can be cut or molded into [different shapes]$_{new}$.]$_{RHEME}$

Figure 1: Annotation of Pie-in-the-Sky sentences with Information Structure

6 Pie-in-the-Sky annotation

"Pie in the Sky" is a joint effort to annotate two sentences with as much semantic/pragmatic information as possible (see `http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html`). Information structure is one of the desired annotation layers. And, as standards are not yet established, our proposal contributes to defining annotation guidelines for this structure. Figure 1 report the Pie-in-the-sky sentences enriched with our annotation. The context prior to these sentences is as follows:

> "a 12-year-old boy reports seeing a man launch a rubber boat from a car parked at the harbor. fbi officials find what they believe may be explosives in the car. yemeni police trace the car to a nearby house. the fbi finds traces of explosives on clothes found neighbors say they saw two men who they describe as "arab-looking" living there for several weeks. police also find a second house where authorities believe two others may have assembled the bomb, possibly doing some welding. passports found in one of the houses identify the men as from a privilege convenience province noted for lawless tribes. but the documents turn out to be fakes. meantime, analysts at the fbi crime lab try to discover what the bomb was made from. no conclusions yet, u.s. officials say. but a working theory, plastic explosive."

We identified 14 NPs markable for information status (see Figure 1).[11] Most annotations were straightforward. Some comments though: "Yemen" is annotated as med/general, although it could also be med/sit as "Yemeni" was previously mentioned. Our decision tree was used for such cases. "The explosive material" is med/set not old/identity since it refers to the kind of explosive used rather than to a specific entity previously mentioned.

In the absence of any prosodic annotation in the transcript, these sentences are slightly ambiguous as to information structure. The most likely interpretation is given in Figure 1.[12] For example, "Yemen's President" contrasts with "US officials", in the set of people talking about what the bomb is made of. Since both words are contrastive, either or both could have L+H* accents, whereas "say" could not. The inclusion of the latter in the theme is consistent with the possibility of a rising boundary LH% after it. "The FBI has told him" is thematic because it links "Yemen's president"'s opinion to the previous discourse. It also would sound appropriate with an L+H*LH% tune. As can be seen, although theme/rheme and prosodic phrase boundaries align, in both cases the VP is split between information/intonation phrases. The independence of information structure and intonation structure from traditional surface structure is a major reason behind our use of 'stand-off' markup.

[11] Square brackets are used to mark annotation boundaries.

[12] Kontrast is marked with the relevant category, unmarked words are background.

7 Applications and Future Work

Once completed, the annotations we have presented, along with those existing for syntax, disfluencies and dialog-acts on the same portion of *Switchboard*, will create a corpus of conversational speech unique in terms of size and richness of annotation. In conjunction with the NXT tools, this resource would optimally lend itself to detailed and rich analysis of diverse linguistic phenomena, the ultimate goal of the Pie in the Sky project. It will be useful for a large range of NLP applications, including paraphrase analysis and generation, topic detection, information extraction and speech synthesis in dialogue systems.

Website Example sound files available at *http://homepages.inf.ed.ac.uk/s0199920/pieinsky.html*.

Acknowledgements Part of this work was funded by Scottish Enterprise (The Edinburgh-Stanford Link *Paraphrase Analysis for Improved Generation* and *Sounds of Discourse*). We would like to thank David Beaver, Jean Carletta, Shipra Dingare, Florian Jaeger, Dan Jurafsky, Vasilis Karaiskos and Bob Ladd for valuable help and discussion.

References

G. M. Ayers. 1994. Discourse functions of pitch range in spontaneous and read speech. In J. Venditti, editor, *OSU Working Papers in Linguistics*, volume 44, pages 1–49.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In C. Boitet and P. Whitelock, editors, *Proc. COLING-ACL*, pages 86–90.

E. Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

S. Baumann, C. Brinckmann, S. Hansen-Schirra, G-J. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proc. NAACL/HLT "Frontiers in Corpus Annotation"*, Boston, MA.

M. Beckman and G. Elam. 1997. Guidelines for ToBI Labelling.. The OSU Research Foundation, v.3.0.

P. Boersma and D. Weenink. 2003. Praat:doing phonetics by computer. http://www.praat.org.

J. M. Brenier, D. M. Cer, and D. Jurafsky. 2005. Emphasis detection in speech using acoustic and lexical features. In *LSA Annual Meeting*, Oakland, CA.

S. Calhoun. 2004a. Overloaded ToBI and what to do about it: An argument for function-based phonological intonation categories. In *Univ. of Edinburgh Ling. Postgrad. Conf.*.

S. Calhoun. 2004b. Phonetic dimensions of intonational categories - L+H* and H*. In *Prosody 2004*, Nara, Japan.

J. Carletta, S. Dingare, M. Nissim, and T. Nikitina. 2004. Using the NITE XML Toolkit on the Switchboard Corpus to study syntactic choice: a case study. In *Proc. of LREC2004, Lisbon.*

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comp. Ling.*, 22(2):249–254.

H. H. Clark. 1975. Bridging. In R. Schank and B. Nash-Webber, eds, *Theoretical Issues in NLP*. MIT Press, Cambridge, MA.

A. Cutler, D. Dahan, and W. van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Lang. and Sp.*, 40(2):141–201.

S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. 1998. Annotating coreference in dialogues: Proposal for a scheme for MATE, `http://www.hcrc.ed.ac.uk/~poesio/anno_manual.html`.

G. Dogil, J. Kuhn, J. Mayer, G. Mhler, and S. Rapp. 1997. Prosody and discourse structure: Issues and experiments. In *Proc. of the ESCA Workshop on Intonation: Theory, Models and Applications*, pages 99–102, Athens, Greece.

M. Eckert and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *J. of Semantics*, 17(1):51–89.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP-92*, pages 517–520.

N. Hedberg and JM. Sosa. 2001. The prosodic structure of topic and focus in spontaneous english dialogue. In *LSA Workshop on Topic and Focus*, Santa Barbara.

M. Heldner, E. Strangert, and T. Deschamps. 1999. A focus detector using overall intensity and high frequency emphasis. In *Proc. ICPhS-99*, vol 2, 1491–1493, San Francisco.

J. Hirschberg. 1993. Pitch accent in context: Predicting intonational prominence from text. *AI*, 63:305–340.

L. Hirschman and N. Chinchor. 1997. MUC-7 coreference task definition. In *Proc. of 7^{th} Conf. on Message Understanding*.

D. R. Ladd. 1996. *Intonational Phonology*. CUP, UK.

K. Lambrecht. 1994. *Information structure and sentence form. Topic, focus, and the mental representation of discourse referents*. Camb. U. Press, UK.

S. Löbner. 1985. Definites. *J. of Semantics*, 4:279–326.

M. Marcus, B. Santorini, and MA. Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. *Comp. Ling.*, 19:313–330.

A. Mengel and W. Lezius. 2000. An XML-based encoding format for syntactically annotated corpora. In *Proc. LREC2000*,121–126.

M. Nissim. 2003. Annotation scheme for information status in dialogue. HCRC, University of Edinburgh. Unpub. ms.

M. Nissim, S. Dingare, J. Carletta, and M. Steedman. 2004. An annotation scheme for information status in dialogue. In *Proc. LREC2004, Lisbon.*

R. Passonneau. 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpub. ms..

J. Pitrelli, M. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labelling reliability in the ToBI framework. In *Proc. of the 3^{rd} Intl. Conf. on Spoken Lge. Proc.*, vol. 2, pages 123–126.

M. Poesio. 2000. The GNOME annotation scheme manual (v.4), `http://www.hcrc.ed.ac.uk/~gnome/anno_manual.html`.

E. F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*. Acad. Press, NY.

E. Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, eds., *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, Philadelphia/Amsterdam.

H.H. Rump and R. Collier. 1996. Focus conditions and the prominence of pitch-accented syllables. *Lang. and Sp.*, 39:1–17.

E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, and C.V. Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. and Sp.*, 41(3-4):439–487.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Sp. Comm.*, 32(2):127–154.

M. Steedman. 2000. Information Structure and the Syntax-Phonology Interface. *LI*, 31(4):649-689.

H. Truckenbrodt. 2002. Upstep and embedded register levels. *Phonology*, 19:77–120.

E. Vallduví and M. Vilkuna. 1998. On Rheme and Kontrast. *Syntax and Semantics*, 29:79-108.

Annotating Attributions and Private States

Theresa Wilson
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
twilson@cs.pitt.edu

Janyce Wiebe
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
wiebe@cs.pitt.edu

Abstract

This paper describes extensions to a corpus annotation scheme for the manual annotation of attributions, as well as opinions, emotions, sentiments, speculations, evaluations and other *private states* in language. It discusses the scheme with respect to the "Pie in the Sky" *Check List of Desirable Semantic Information for Annotation*. We believe that the scheme is a good foundation for adding private state annotations to other layers of semantic meaning.

1 Introduction

This paper describes a fine-grained annotation scheme for key components and properties of opinions, emotions, sentiments, speculations, evaluations, and other *private states* in text. We first give an overview of the core scheme. We then describe recent extensions to the scheme, namely refined annotations of *attitudes* and *targets*, or objects, of private states. Finally, we discuss related items from the "Pie in the Sky" *Check List of Desirable Semantic Information for Annotation*, and related work. We believe our scheme would provide a foundation for adding private state annotations to other layers of semantic and pragmatic meaning.

2 The Core Scheme

This section overviews the core of the annotation scheme. Further details may be found in (Wilson and Wiebe, 2003; Wiebe et al., 2005).

2.1 Means of Expressing Private States

The goals of the annotation scheme are to represent internal mental and emotional states, and to distinguish subjective information from material presented as fact. As a result, the annotation scheme is centered on the notion of *private state*, a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. As Quirk et al. (1985) define it, a *private state* is a state that is not open to objective observation or verification: "a person may be observed to *assert that God exists*, but not to *believe that God exists*. Belief is in this sense 'private'." (p. 1181) Following literary theorists such as Banfield (1982), we use the term *subjectivity* for linguistic expressions of private states in the contexts of texts and conversations.

We can further view private states in terms of their functional components — as states of *experiencers* holding *attitudes*, optionally toward *targets*. For example, for the private state in the sentence *John hates Mary*, the experiencer is "John," the attitude is "hate," and the target is "Mary."

We create private state frames for three main types of private state expressions in text:

- explicit mentions of private states
- speech events expressing private states
- expressive subjective elements

An example of an explicit mention of a private state is "fears" in (1):

> (1) "The U.S. **fears** a spill-over," said Xirao-Nima.

An example of a *speech event* expressing a private state is "said" in (2):

> (2) "The report is **full of absurdities**," Xirao-Nima **said**.

Note that we use the term *speech event* to refer to both speaking and writing events.

The phrase "full of absurdities" in (2) above is an *expressive subjective element* (Banfield, 1982). Other examples can be found in (3):

> (3) **The time has come**, **gentlemen**, for Sharon, **the assassin**, to realize that **injustice cannot last long**.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 53–60,
Ann Arbor, June 2005.

The private states in this sentence are expressed entirely by the words and the style of language that is used. In (3), although the writer does not explicitly say that he hates Sharon, his choice of words clearly demonstrates a negative attitude toward him. As used in these sentences, the phrases "The time has come," "gentlemen," "the assassin," and "injustice cannot last long" are all expressive subjective elements. Expressive subjective elements are used by people to express their frustration, anger, wonder, positive sentiment, etc., without explicitly stating that they are frustrated, angry, etc. Sarcasm and irony often involve expressive subjective elements.

2.2 Private State Frames

We propose two types of private state frames: *expressive subjective element frames* will be used to represent expressive subjective elements; and *direct subjective frames* will be used to represent both subjective speech events (i.e., speech events expressing private states) and explicitly mentioned private states. The frames have the following attributes:

Direct subjective (subjective speech event or explicit private state) frame:

- **text anchor:** a pointer to the span of text that represents the speech event or explicit mention of a private state.
- **source:** the person or entity that expresses or experiences the private state, possibly the writer.
- **target:** the target or topic of the private state, i.e., what the speech event or private state is about.
- **properties:**
  - **intensity:** the intensity of the private state (*low*, *medium*, *high*, or *extreme*).
  - **expression intensity:** the contribution of the speech event or private state expression itself to the overall intensity of the private state. For example, "say" is often neutral, even if what is uttered is not neutral, while "excoriate" itself implies a very strong private state.
  - **insubstantial:** true, if the private state is not substantial in the discourse. For example, a private state in the context of a conditional often has the value *true* for attribute *insubstantial*.
  - **attitude type:** the type of attitude(s) composing the private state.

Expressive subjective element frame:

- **text anchor:** a pointer to the span of text that denotes the subjective or expressive phrase.
- **source:** the person or entity that is expressing the private state, possibly the writer.
- **properties:**
  - **intensity:** the intensity of the private state.
  - **attitude type**

2.3 Objective Speech Event Frames

To distinguish opinion-oriented material from material presented as factual, we also define *objective speech event frames*. These are used to represent material that is attributed to some source, but is presented as objective fact. They include a subset of the slots in private state frames:

Objective speech event frame:

- **text anchor:** a pointer to the span of text that denotes the speech event.
- **source:** the speaker or writer.
- **target:** the target or topic of the speech event, i.e., the content of what is said.

For example, an objective speech event frame is created for "said" in the following sentence (assuming no undue influence from the context):

> (4) Sargeant O'Leary said the incident took place at 2:00pm.

That the incident took place at 2:00pm is presented as a fact with Sargeant O'Leary as the source of information.

2.4 Agent Frames

The annotation scheme includes an *agent frame* for noun phrases that refer to sources of private states and speech events, i.e., for all noun phrases that act as the experiencer of a private state, or the speaker/writer of a speech event. Each agent frame generally has two slots. The *text anchor* slot includes a pointer to the span of text that denotes the noun phrase source. The *source* slot contains a unique alpha-numeric ID that is used to denote this source throughout the document. The agent frame associated with the first informative (e.g., non-pronominal) reference to this source in the document includes an *id* slot to set up the document-specific source-id mapping.

2.5 Nested Sources

The source of a speech event is the speaker or writer. The source of a private state is the experiencer of the private state, i.e., the person whose opinion or emotion is being expressed. The writer of an article is always a source, because he or she wrote the sentences of the article, but the writer may also write about other people's private states

and speech events, leading to multiple sources in a single sentence. For example, each of the following sentences has two sources: the writer (because he or she wrote the sentences), and Sue (because she is the source of a speech event in (5) and of private states in (6) and (7)).

(5) Sue said, "The election was fair."
(6) Sue thinks that the election was fair.
(7) Sue is afraid to go outside.

Note, however, that we don't really know what Sue says, thinks or feels. All we know is what the writer tells us. For example, Sentence (5) does not directly present Sue's speech event but rather Sue's speech event according to the writer. Thus, we have a natural *nesting of sources* in a sentence.

In particular, private states are often filtered through the "eyes" of another source, and private states are often directed toward the private states of others. Consider sentence (1) above and (8) following:

(8) China criticized the U.S. report's criticism of China's human rights record.

In sentence (1), the U.S. does not directly state its fear. Rather, according to the writer, according to Xirao-Nima, the U.S. fears a spill-over. The source of the private state expressed by "fears" is thus the *nested source* ⟨*writer, Xirao-Nima, U.S.*⟩. In sentence (8), the U.S. report's criticism is the target of China's criticism. Thus, the nested source for "criticism" is ⟨*writer, China, U.S. report*⟩.

Note that the shallowest (left-most) agent of all nested sources is the writer, since he or she wrote the sentence. In addition, nested source annotations are composed of the IDs associated with each source, as described in the previous subsection. Thus, for example, the nested source ⟨*writer, China, U.S. report*⟩ would be represented using the IDs associated with the writer, China, and the report being referred to, respectively.

2.6 Examples

We end this section with examples of direct subjective, expressive subjective element, and objective speech event frames (sans target and attitude type attributes, which are discussed in the next section).

First, we show the frames that would be associated with sentence (9), assuming that the relevant source ID's have already been defined:

(9) "The US fears a spill-over," said Xirao-Nima.

```
Objective speech event:
   Text anchor: the entire sentence
   Source: <writer>
   Implicit: true

Objective speech event:
   Text anchor: said
   Source: <writer,Xirao-Nima>

Direct subjective:
   Text anchor: fears
   Source: <writer,Xirao-Nima,U.S.>
   Intensity: medium
   Expression intensity: medium
```

The first objective speech event frame represents that, according to the writer, it is true that Xirao-Nima uttered the quote and is a professor at the university referred to. The *implicit* attribute is included because the writer's speech event is not explicitly mentioned in the sentence (i.e., there is no explicit phrase such as "I write").

The second objective speech event frame represents that, according to the writer, according to Xirao-Nima, it is true that the US fears a spillover. Finally, when we drill down to the subordinate clause we find a private state: the US fear of a spillover. Such detailed analyses, encoded as annotations on the input text, would enable a person or an automated system to pinpoint the subjectivity in a sentence, and attribute it appropriately.

Now, consider sentence (10):

(10) "The report is full of absurdities," Xirao-Nima said.

```
Objective speech event:
   Text anchor: the entire sentence
   Source: <writer>
   Implicit: true

Direct subjective:
   Text anchor: said
   Source: <writer,Xirao-Nima>
   Intensity: high
   Expression intensity: neutral

Expressive subjective element:
   Text anchor: full of absurdities
   Source: <writer,Xirao-Nima>
   Intensity: high
```

The objective frame represents that, according to the writer, it is true that Xirao-Nima uttered the quoted string. The second frame is created for "said" because it is a subjective speech event: private states are conveyed in what is uttered. Note that *intensity* is *high* but *expression intensity* is *neutral*: the private state being expressed is strong, but the specific speech event phrase "said" does not itself contribute to the intensity of the private state. The third frame is for the expressive subjective element "full of absurdities."

3 Annotation Process

To date, over 11,000 sentences in 550 documents have been annotated according to the annotation scheme described above. The documents are English-language versions of news documents from the world press. The documents are from 187 different news sources in a variety

of countries. The original documents and their annotations are available at
http://nrrc.mitre.org/NRRC/publications.htm.

The annotation process and inter-annotator agreement studies are described in (Wiebe et al., 2005). Here, we want to highlight two themes of the annotation instructions:

1. There are no fixed rules about how particular words should be annotated. The instructions describe the annotations of specific examples, but do not state that specific words should always be annotated a certain way.

2. Sentences should be interpreted with respect to the contexts in which they appear. The annotators should not take sentences out of context and think what they *could* mean, but rather should judge them as they are being used in that particular sentence and document.

We believe that these general strategies for annotation support the creation of corpora that will be useful for studying expressions of subjectivity in context.

4 Extensions: Attitude and Target Annotations

Before we describe the new attitude and target annotations, consider the following sentence.

> (11) "I think people are happy because Chavez has fallen."

This sentence contains two private states, represented by direct subjective annotations anchored on "think" and "happy," respectively.

The word "think" is used to express an opinion about what is true according to its source (a *positive arguing* attitude type; see Section 4.1). The target of "think" is "people are happy because Chavez has fallen."

The word "happy" clearly expresses a positive attitude, with target "Chavez has fallen." However, looking more closely at the private state for "happy," we see that we can also infer a negative attitude toward Chavez, from the phrase "happy because Chavez has fallen."

Sentence (11) illustrates some of the things we need to consider when representing attitudes and targets. First, we see that more than one type of attitude may be involved when a private state is expressed. In (11), there are three (a positive attitude, a negative attitude, and a positive arguing attitude). Second, more than one target may be associated with a private state. Consider "happy" in (11). The target of the positive attitude is "Chavez has fallen," while the target of the inferred negative attitude is "Chavez."

Positive Attitudes	Positive Arguing
Negative Attitudes	Negative Arguing
Positive Intentions	Speculation
Negative Intentions	Other Attitudes

Table 1: Attitude Types

The representation also must support multiple targets for a single attitude, as illustrated by Sentence (12):

> (12) Tsvangirai said the election result was a clear case of highway robbery by Mugabe, his government and his party, Zanu-PF.

In (12), the phrase "a clear case of highway robbery" expresses a negative attitude of Tsvangirai. This negative attitude has two targets: "the election results" and "Mugabe, his government and his party, Zanu-PF."

To capture the kind of detailed attitude and target information that we described above, we propose two new types of annotations: *attitude frames* and *target frames*. We describe these new annotations in Sections 4.2 and 4.3, but first we introduce the set of attitude types that we developed for the annotation scheme.

4.1 Types of Attitudes

One of our goals in extending the annotation scheme for private states was to develop a set of attitude types that would be useful for NLP applications. It it also important that the set of attitude types provide good coverage for the range of possible private states. Working with our annotators and looking at the private states already annotated, we developed the set of attitude types listed in Table 1.

Below we give a brief description of each attitude type, followed by an example. In each example, the span of text that expresses the attitude type is in bold, and the span of text that refers to the target of the attitude type (if a target is given) is in angle brackets.

Positive Attitudes: positive emotions, evaluations, judgments and stances.

> (13) The Namibians went as far as to say ⟨Zimbabwe's election system⟩ was "**water tight, without room for rigging**".

Negative Attitudes: negative emotions, evaluations, judgments and stances.

> (14) His disenfranchised supporters **were seething**.

Positive Arguing: arguing for something, arguing that something is true or so, arguing that something did happen or will happen, etc.

(15) Iran **insists** ⟨its nuclear program is purely for peaceful purposes⟩.

Negative Arguing: arguing against something, arguing that something is not true or not so, arguing that something did not happen or will not happen, etc.

(16) Officials in Panama **denied** that ⟨Mr. Chavez or any of his family members had asked for asylum⟩.

Positive Intentions: aims, goals, plans, and other overtly expressed intentions.

(17) The Republic of China government believes in the US **committment** ⟨to separating its anti-terrorism campaign from the Taiwan Strait issue⟩, an official said Thursday.

Negative Intentions: expressing that something is not an aim, not a goal, not an intention, etc.

(18) The Bush administration **has no plans** ⟨to ease sanctions against mainland China⟩.

Speculation: speculation or uncertainty about what may or may not be true, what may or may not happen, etc.

(19) ⟨The president is **likely** to endorse the bill⟩.

Other Attitudes: other types of attitudes that do not fall into one of the above categories.

(20) To the **surprise** of many, ⟨the dollar hit only 2.4 pesos and closed at 2.1⟩.

4.2 Attitude Frames

With the introduction of the attitude frames, two issues arise. First, which spans of text should the new attitudes be anchored to? Second, how do we tie the attitude frames back to the private states that they are part of?

The following sentence illustrates the first issue.

(21) The MDC leader said systematic cheating, spoiling tactics, rigid new laws, and shear obstruction - as well as political violence and intimidation - were just some of the irregularities practised by the authorities in the run-up to, and during the poll.

In (21), there are 5 private state frames attributed to the MDC leader: a direct subjective frame anchored to "said," and four expressive subjective element frames anchored respectively to "systematic cheating . . . obstruction," "as well as," "violence and intimidation," and "just some of the irregularities." We could create an attitude frame for each of these private state frames, but we believe the following is a better solution. For each direct subjective frame, the annotator is asked to consider the direct subjective annotation and everything within the scope of the annotation when deciding what attitude types are being expressed by the source of the direct subjective frame. Then, for each attitude type identified, the annotator creates an attitude frame and anchors the frame to whatever span of text completely captures the attitude type. In to sentence (21), this results in just one attitude frame being created to represent the negative attitude of the MDC leader. The anchor for this attitude frame begins with "systematic cheating" and ends with "irregularities."

Turning to the second issue, tying attitude frames to their private states, we do two things. First, we create a unique ID for the attitude frame. Then, we change the attitude type attribute on the direct subjective annotation into a new attribute called an *attitude link*. We place the attitude frame ID into the attitude link slot. The attitude link slot can hold more then one attitude frame ID, allowing us to represent a private state composed of more than one type of attitude.

Because we expect the attitude annotations to overlap with most of the expressive subjective element annotations, we chose not to link attitude frames to expressive subjective element frames. However, this would be possible to do should it become necessary.

The attitude frame has the following attributes:

Attitude frame:

- **id:** a unique alphanumeric ID for identifying the attitude annotation. The ID is used to link the attitude annotation to the private state it is part of.
- **text anchor:** a pointer to the span of text that captures the attitude being expressed.
- **attitude type:** one of the attitude types listed in Table 1.
- **target link:** one or more target annotation IDs (see Section 4.3).
- **intensity:** the intensity of the attitude.
- **properties:**
  - **inferred:** true, if the attitude is inferred.
  - **sarcastic:** true, if the attitude is realized through sarcasm.
  - **repetition:** true, if the attitude is realized through the repetition of words, phrases, or syntax.
  - **contrast:** true, if the attitude is realized only through contrast with another attitude.

Of the four attitude-frame properties, *inferred* was already discussed. The property *sarcastic* marks attitudes expressed using sarcasm. In general, we think this property will be of interest for NLP applications working with opinions. Detecting sarcasm may also help a system learn to distinguish between positive and negative attitudes. The sarcasm in Sentence (22), below, makes the word "Great" an expression of a negative rather than a positive attitude.

> (22) "Great, keep on buying dollars so there'll be more and more poor people in the country," shouted one.

The *repetition* and *contrast* properties are also for marking different ways in which an attitude might be realized. We feel these properties will be useful for developing an automatic system for recognizing different types of attitudes.

4.3 Target Frames

The target frame is used to mark the target of each attitude. A target frame has two slots, the *id* slot and the *text anchor* slot. The id slot contains a unique alpha-numeric ID for identifying the target annotation. We use the target frame ID to link the target back to the attitude frame. The attitude frame has a *target-link* slot that can hold one or more target frame IDs. This allows us to represent when a single attitude is directed at more than one target.

The text anchor slot has a pointer to the span of text that denotes the target. If there is more than one reference to the target in the sentence, the most *syntactically relevant* reference is chosen.

To illustrate what we mean by syntactically relevant, consider the following sentence.

> (23) African observers **generally approved** of ⟨his victory⟩ while Western governments **denounced** ⟨it⟩.

The target of the two attitudes (in bold) in the above sentence is the same entity in the discourse. However, although we anchor the target for the first attitude to "his victory," the anchor for the target of the second attitude is the pronoun "it." As the direct object of the span that denotes the attitude "denounced," "it" is more syntactically relevant than "his victory."

4.4 Illustrative Examples

Figures 4.4 and 4.4 give graphical representations for the annotations in sentences (11) and (12). With attitude frame and target frame extensions, we are able to capture more detail about the private states being expressed in the text than the original core scheme presented in (Wiebe et al., 2005).

5 Pie in the Sky Annotation

Among the items on the "Pie in the Sky" *Check List of Desirable Semantic Information for Annotation*, [1] the most closely related are *epistemic values ("attitude?")*, *epistemic*, *deontic*, and *personal attitudes*. These all fundamentally involve a *self* (Banfield, 1982), a subject of consciousness who is the source of knowledge assessments, judgments of certainty, judgments of obligation/permission, personal attitudes, and so on. Any explicit epistemic, deontic, or personal attitude expressions are represented by us as private state frames, either direct subjective frames (e.g., for verbs such as "know" referring to an epistemic state) or expressive subjective element frames (e.g., for modals such as "must" or "ought to"). Importantly, many deontic, epistemic, and personal attitude expressions do not directly express the speaker or writer's subjectivity, but are attributed by the speaker or writer to agents mentioned in the text (consider, e.g., "John believes that Mary should quit her job"). Our frame and nested-source representations were designed to support attributing subjectivity to appropriate sources. In future work, additional attributes could be added to private state frames to distinguish between, for example, deontic and epistemic usages of "must" and to represent different epistemic values.

Other phenomena on the list overlap with subjectivity, such as *modality* and *social style/register*. As mentioned above, some modal expressions are subjective, such as those expressing deontic or epistemic judgments. However, hypotheticals and future expressions need not be subjective. For example, "The company announced that if its profits decrease in the next quarter, it will lay off some employees" may easily be interpreted as presenting objective fact. As for style, some are subjective by their nature. One is the literary style *represented thought*, used to present consciousness in fiction (Cohn, 1978; Banfield, 1982). Others are sarcastic or dismissive styles of speaking or writing. In our annotation scheme, sentences perceived to represent a character's consciousness are represented with private-state frames, as are expressions perceived to be sarcastic or dismissive. On the other hand, some style distinctions, such as degree of formality, are often realized in other ways than with explicit subjective expressions (e.g., "can't" versus "cannot").

Polarity, another item on the checklist, also overlaps with subjective positive and negative attitude types. Although many negative and positive polarity words are seldom used outside subjective expressions (such as "hate" and "love"), others often are. For example, words such as "addicted" and "abandoned" are included as negative polarity terms in the General Inquirer lexicon (General-Inquirer, 2000), but they can easily appear in objective

[1] Available at: http://nlp.cs.nyu.edu/meyers/frontiers/2005.html

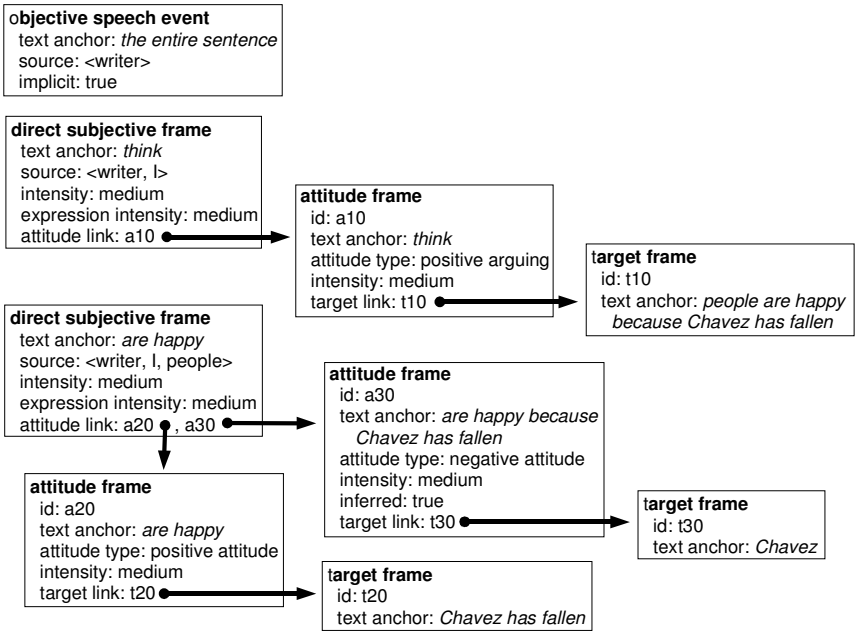


Figure 1: Graphical representation of annotations for Sentence (11)

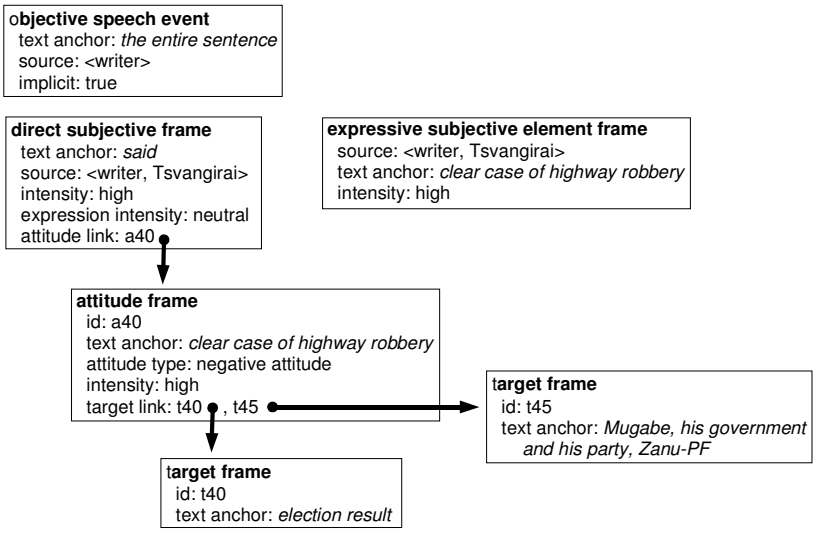


Figure 2: Graphical representation of annotations for Sentence (12)

sentences (e.g., "Thomas De Quincy was addicted to opium and lived in an abandoned shack").

Integrating subjectivity with other layers of annotation proposed in the "Pie in the Sky" project would afford the opportunity to investigate how they interact. It would also enrich our subjectivity representations. While our scheme promises to be a good base, much remains to be added. For example, annotations of thematic roles and co-reference would add needed structure to the target annotations, which are now only spans of text. In addition, temporal and modal annotations would flesh out the *insubstantial* attribute, which is currently only a binary marker. Furthermore, individual private state expressions must be integrated with respect to the discourse context. For example, which expressions of opinions oppose versus support one another? Which sentences presented as objective fact are included to support a subjective opinion? A challenging dimension to add to the "Pie in the Sky" project would be the *deictic center* as conceived of in (Duchan et al., 1995), which consists of *here*, *now*, and *I* reference points updated as the text or conversation unfolds. Our annotation scheme was developed with this framework in mind.

6 Related Work

The work most similar to ours is Appraisal Theory (Martin, 2000; White, 2002) from systemic functional linguistics (see Halliday (19851994)). Both Appraisal Theory and our annotation scheme are concerned with identifying and characterizing expressions of opinions and emotions in context. The two schemes, however, make different distinctions. Appraisal Theory distinguishes different types of positive and negative attitudes and also various types of "intersubjective positioning" such as attribution and expectation. Appraisal Theory does not distinguish, as we do, the different ways that private states may be expressed (i.e., directly, or indirectly using expressive subjective elements). It also does not include a representation for nested levels of attribution.

In addition to Appraisal Theory, subjectivity annotation of text in context has also been performed in Yu and Hatzivassiloglou (2003), Bruce and Wiebe (1999), and Wiebe et al. (2004). The annotations in Yu and Hatzivassiloglou (2003) are sentence-level subjective vs. objective and polarity judgments. The annotation schemes used in Bruce and Wiebe (1999) and Wiebe et al. (2004) are earlier, much less detailed versions of the annotation scheme presented in this paper.

7 Conclusion

We have described extensions to an annotation scheme for private states and objective speech events in language. We look forward to integrating and elaborating this scheme with other layers of semantic meaning in the future.

8 Acknowledgments

This work was supported in part by the National Science Foundation under grant IIS-0208798 and by the Advanced Research and Development Activity (ARDA).

References

A. Banfield. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

R. Bruce and J. Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

D. Cohn. 1978. *Transparent Minds: Narrative Modes for Representing Consciousness in Fiction*. Princeton University Press, Princeton, NJ.

J. Duchan, G. Bruder, and L. Hewitt, editors. 1995. *Deixis in Narrative: A Cognitive Science Perspective*. Lawrence Erlbaum Associates.

The General-Inquirer. 2000. *http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm*.

M.A.K. Halliday. 1985/1994. *An Introduction to Functional Grammar*. London: Edward Arnold.

J.R. Martin. 2000. Beyond exchange: APPRAISAL systems in English. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text: Authorial stance and the construction of discourse*, pages 142–175. Oxford: Oxford University Press.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

P.R.R. White. 2002. Appraisal: The language of attitudinal evaluation and intersubjective stance. In Verschueren, Ostman, blommaert, and Bulcaen, editors, *The Handbook of Pragmatics*, pages 1–27. Amsterdam/Philadelphia: John Benjamins Publishing Company.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evalution (formerly Computers and the Humanities)*, 1(2).

T. Wilson and J. Wiebe. 2003. Annotating opinions in the world press. In *SIGdial-03*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP-2003*.

A Parallel Proposition Bank II for Chinese and English*

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, Benjamin Snyder
Department of Computer and Information Science
University of Pennsylvania
{mpalmer/xueniwen/malayao/Jinying/bsnyder3}@linc.cis.upenn.edu

Abstract

The Proposition Bank (PropBank) project is aimed at creating a corpus of text annotated with information about semantic propositions. The second phase of the project, PropBank II adds additional levels of semantic annotation which include eventuality variables, co-reference, coarse-grained sense tags, and discourse connectives. This paper presents the results of the parallel PropBank II project, which adds these richer layers of semantic annotation to the first 100K of the Chinese Treebank and its English translation. Our preliminary analysis supports the hypothesis that this additional annotation reconciles many of the surface differences between the two languages.

1 Introduction

There is a pressing need for a consensus on a task-oriented level of semantic representation that can enable the development of powerful new semantic analyzers in the same way that the Penn Treebank (Marcus et al., 1993) enabled the development of statistical syntactic parsers (Collins, 1999; Charniak, 2001). We believe that shallow semantics expressed as a dependency structure, i.e., predicate-argument structure, for verbs, participial modifiers, and nominalizations provides a feasible level of annotation that would be of great benefit. This annotation, coupled with word senses, minimal co-reference links, event identifiers, and discourse and temporal relations, could provide the foundation for a major advance in our ability to automatically extract salient relationships from text. This will in turn facilitate breakthroughs in message understanding, machine translation, fact retrieval, and information retrieval. The Proposition Bank project is a major step towards providing this type of annotation. It takes a practical approach to semantic representation, adding a layer of predicate argument information, or semantic roles, to the syntactic structures of the Penn Treebank (Palmer et al., 2005). The Frame Files that provide guidance to the annotators constitute a rich English lexicon with explicit ties between syntactic realizations and coarse-grained senses, Framesets. PropBank Framesets are distinguished primarily by syntactic criteria such as differences in subcategorization frames, and can be seen as the top-level of an hierarchy of sense distinctions. Groupings of fine-grained WordNet senses, such as those developed for Senseval2 (Palmer et al., to appear) provide an intermediate level, where groups are distinguished by either syntactic or semantic criteria. WordNet senses constitute the bottom level. The PropBank Frameset distinctions, which can be made consistently by humans and systems (over 90% accuracy for both), are surprisingly compatible with the groupings; 95% of the groups map directly onto a single PropBank frameset sense (Palmer et al., 2004).

The semantic annotation provided by PropBank is only a first approximation at capturing the full richness of semantic representation. Additional annotation of nominalizations and other noun pred-

*This work is funded by the NSF via Grant EIA02-05448 .

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 61–67,
Ann Arbor, June 2005.

icates has already begun at NYU. This paper describes the results of PropBank II, a project to provide richer semantic annotation to structures that have already been propbanked, specifically, eventuality ID' s, coreference, coarse-grained sense tags, and discourse connectives. Of special interest to the machine translation community is our finding, presented in this paper, that PropBank II annotation reconciles many of the surface differences of the two languages.

2 PropBank I

PropBank (Palmer et al., 2005) is an annotation of the Wall Street Journal portion of the Penn Treebank II (Marcus et al., 1994) with 'predicate-argument' structures, using sense tags for highly polysemous words and semantic role labels for each argument. An important goal is to provide consistent semantic role labels across different syntactic realizations of the same verb, as in *the window* in *[ARG0 John] broke [ARG1 the window]* and *[ARG1 The window] broke*. PropBank can provide frequency counts for (statistical) analysis or generation components in a machine translation system, but provides only a shallow semantic analysis in that the annotation is close to the syntactic structure and each verb is its own predicate.

In PropBank, semantic roles are defined on a verb-by-verb basis. An individual verb's semantic arguments are simply numbered, beginning with 0. Polysemous verbs have several *framesets*, corresponding to a relatively coarse notion of word senses, with a separate set of numbered roles, a roleset, defined for each Frameset. For instance, *leave* has both a DEPART Frameset *([ARG0 John] left [ARG1 the room])* and a GIVE Frameset, *([ARG0 I] left [ARG1 my pearls] [ARG2 to my daughter-in-law] [ARGM-LOC in my will].)* While most Framesets have three or four numbered roles, as many as six can appear, in particular for certain verbs of motion. Verbs can take any of a set of general, adjunct-like arguments (ARGMs), such as LOC (location), TMP (time), DIS (discourse connectives), PRP (purpose) or DIR (direction). Negations (NEG) and modals (MOD) are also marked.

There are several other annotation projects, FrameNet (Baker et al., 1998), Salsa (Ellsworth et al., 2004), and the Prague Tectogrammatics (Hajicova and Kucerova, 2002), that share similar goals. Berkeley' s FrameNet project, (Baker et al., 1998; Fillmore and Atkins, 1998; Johnson et al., 2002) is committed to producing rich semantic frames on which the annotation is based, but it is less concerned with annotating complete texts, concentrating instead on annotating a set of examples for each predicator (including verbs, nouns and adjectives), and attempting to describe the network of relations among the semantic frames. For instance, the *buyer* of a *buy* event and the *seller* of a *sell* event would both be Arg0' s (Agents) in PropBank, while in FrameNet one is the BUYER and the other is the SELLER. The Salsa project (Ellsworth et al., 2004) in Germany is producing a German lexicon based on the FrameNet semantic frames and annotating a large German newswire corpus. PropBank style annotation is being used for verbs which do not yet have FrameNet frames defined.

The PropBank annotation philosophy has been extended to the Penn Chinese Proposition Bank (Xue and Palmer, 2003). The Chinese PropBank annotation is performed on a smaller (250k words) and yet growing corpus annotated with syntactic structures (Xue et al., To appear). The same syntactic alternations that form the basis for the English PropBank annotation also exist in robust quantities in Chinese, even though it may not be the case that the same exact verbs (meaning verbs that are close translations of one another) have the exact same range of syntactic realization for Chinese and English. For example, in (1), "新年/New Year 招待会/reception" plays the same role in (a) and (b), which is the event or activity held, even though it occurs in different syntactic positions. Assigning the same argument label, Arg1, to both instances, captures this regularity. It is worth noting that the predicate "举行/hold" does not have passive morphology in (1a), despite what its English translation suggests. Like the English PropBank, the adjunct-like elements receive more general labels like TMP or LOC, as also illustrated in (1). The functional tags for Chinese and English PropBanks are to a large extent similar and more details can be found in (Xue and Palmer, 2003).

(1) a. [ARG1 新年/New Year 招待会/reception] [ARGM-TMP 今 天/today] [ARGM-LOC 在/at 钓 鱼

台/Diaoyutai 国宾馆/state guest house 举行/hold]
"The New Year reception was held in Diao-yutai State Guest House today."

b. [ARG0 唐家旋/Tang Jiaxuan] [ARGM-TMP 今天/today] [ARGM-LOC 在/at 钓鱼台/Diaoyutai 国宾馆/state guest house] 举行/ hold [arg1 新年/New Year 招待会/reception]
"Tang Jiaxuan was holding the New Year reception in Diaoyutai State Guest House today."

3 A Parallel PropBank II

As discussed above, PropBank II adds richer semantic annotation to the PropBank I predicate argument structures, notably eventuality variables, co-references, coarse-grained sense tags (Babko-Malaya et al., 2004; Babko-Malaya and Palmer, 2005), and discourse connectives (Xue, To appear) To create our parallel PropBank II, we began with the first 100K words of the Chinese Treebank which had already been propbanked, and which we had had translated into English. The English translation was first treebanked and then propbanked, and we are now in the process of adding the PropBank II annotation to both the English and the Chinese propbanks. We will discuss our progress on each of the three individual components of PropBank II in turn, bringing out translation issues along the way that have been highlighted by the additional annotation. In general we find that this level of abstraction facilitates the alignment of the source and target language descriptions: event ID' s and event coreferences simplify the mappings between verbal and nominal events; English coarse-grained sense tags correspond to unique Chinese lemmas; and discourse connectives correspond well.

3.1 Eventuality variables

Positing eventuality[1] variables provides a straightforward way to represent the semantics of adverbial modifiers of events and capture nominal and pronominal references to events. Given that the arguments and adjuncts for the verbs are already annotated in Propbank I, adding eventuality variables is for the most part straightforward. The example in (2) illustrates a Propbank I annotation, which is identified with a unique event id in Propbank II.

(2) a. Mr. Bush met him privately in the White House on Thursday.

b. Propbank I: Rel: met, Arg0: Mr. Bush, Arg1: him, ArgM-MNR: privately, ArgM-LOC: in the White House, ArgM-TMP: on Thursday.

c. Propbank II: $\exists e$ meeting(e) & Arg0(e,Mr. Bush) & Arg1(e, him) & MNR (e, privately) & LOC(e, in the White House) & TMP (e, on Thursday).

Annotation of event variables starts by automatically associating all Propbank I annotations with potential event ids. Since not all annotations actually denote eventualities, we manually filter out selected classes of verbs. We further attempt to identify all nouns and nominals which describe eventualities as well as all sentential arguments of the verbs which refer to events. And, finally, part of the PropBank II annotation involves tagging of event coreference for pronouns as well as empty categories. All these tasks are discussed in more detail below.

Identifying event modifiers. The actual annotation starts from the presumption that all verbs are events or states and nouns are not. All the verbs in the corpus are automatically assigned a unique event identifier and the manual part of the task becomes (i) identification of verbs or verb senses that do not denote eventualities, (ii) identification of nouns that do denote events. For example, in (3), *begin* is an aspectual verb that does not introduce an event variable, but rather modifies the verb 'take', as is supported by the fact that it is translated as an adverb "初/initially" in the corresponding Chinese sentence.

(3) 重点/key 发展/develop 的/DE 医药/medicine 与/and 生物/biology 技术/technology, 新/new 技术/technology, 新/new 材料/material, 计算机/computer 及/and 应用/application, 光/photo 电/electric 一体化/integration 等/etc. 产业/industry 已/already 初/initially 具/take 规模/shape.
"Key developments in industries such as medicine, biotechnology, new materials, computer and its applications, protoelectric integration, etc. have begun to take shape."

Nominalizations as events Although most nouns do not introduce eventualities, some do and these nouns are generally nominalizations[2] . This is true

[1]The term 'eventuality' is used here to refer to events and states.

[2]The problem of identifying nouns which denote events is addressed as part of the sense-tagging tagging. Detailed discussion can be found in (Babko-Malaya and Palmer, 2005).

for both English and Chinese, as is illustrated in (4). Both "发展/develop" and "深入/deepening" are nominalized verbs that denote events. Having a parallel propbank annotated with event variables allows us to see how events are lined up in the two languages and how their lexical realizations can vary. The nominalized verbs in Chinese can be translated into verbs or their nominalizations, as is shown in the alternative translations of the Chinese original in (4). What makes this particular example even more interesting is the fact that the adjective modifier of the events, "不断/continued", can actually be realized as an aspectual verb in English. The semantic representations of the Propbank II annotation, however, are preserved: both the aspectual verb "continue" in English and the adjective "不断/continued" in Chinese are modifiers of the events denoted by "发展/development" and "深入/deepening".

(4) 随/with 着/with 中/China 国/China 经/economy 济/economy 的/DE 不断/**continued** 发展/development 和/and 对/to 外/outside 开放/open 的/DE 不断/**continued** 深入/deepen ...
"As China's economy **continues** to develop and its practice of opening to the outside **continues** to deepen..."
"With the continued development of China's economy and the continued deepening of its practice of opening to the outside..."

Event Coreference Another aspect of the event variable annotation involves identifying pronominal expressions that corefer with events. These pronominal expressions may be overt, as in the Chinese example in (5), while others correspond to null pronouns, marked as **pro**[3]. in the Treebank annotations, as in (6):

(5) 而且/additionally, 出口/export 商品/commodity 结构/structure 继续/continue 优化/optimize, 去年/last year 工业/industry 制成品/finished product 出口/export 额/quota 占/account for 全国/entire country 出口/export 总额/quantity 的/DE 比重/proportion 达/reach 百分之八十五点六/85.6 percent, 这/**this** 充分/clearly 表明/indicate 中国/China 工业/industry 产品/product 的/DE 制造/produce 水平/level 比/compared with 过去/past 有/have 了/LE 很/very 大/big 提高/improvement.
"Moreover, the structure of export com-modities continues to optimize, and last year's export volume of manufactured products ac-counts for 85.6 percent of the whole countries' export, ***pro*** clearly indicating that China's industrial product manufacturing level has improved."

(6) 这些/these 成果/achievement 中/among 有/have 一百三十八/138 项/item 被/BEI 企业/enterprise 应用/apply 到/to 生产/production 上/on "点石成金/spin gold from straw", ***pro*** 大大/greatly 提高/improve 了/ASP 中国/China 镍/nickel 工业/industry 的/DE 生产/production 水平/level.
"Among these achievements, 138 items have been applied to production by enterprises to spin gold from straw, which greatly improved the production level of China's nickel industry."

It is not the case, however that overt pro-nouns in Chinese will always correspond to overt pronouns in English. In (5), the overt pronoun "这/this" in Chinese corresponds with a null pronoun in English in the beginning of a reduced relative clause, while in (6), the null pronoun in Chinese is translated into a relative pronoun "which" that introduces a relative clause. In other cases, neither language has an overt pronoun, although one is posited in the treebank annotation, as in (7).

(7) 去年/last year, 纽约/New York 新/new 上市/list 的/DE 外国/foreign 企业/enterprise 共/altogether 有/have 61/61 家/CL, ***pro*** 创/create 历年/recent year 来/since 最高/highest 纪录/record.
"Last year, there were 61 new foreign en-terpises listed in New York Stock Exchange, ***PRO*** creating the highest record in history."

Having a parallel propbank annotated with event variables allows us to examine how the same events are lexicalized in English and Chi-nese and how they align, whether they have been indicated by verbs or nouns.

3.2 Grouped sense tags

In general, the verbs in the Chinese PropBank are less polysemous than the English PropBank verbs, with the vast majority of the lemmas having just one Frameset. On the other hand, the Chinese PropBank has more lemmas (including stative verbs which are generally translated into adjectives in English) normalized by the corpus size. The Chinese PropBank has 4854 lemmas in the 250K words that have been propbanked alone, while the English PropBank has just 3635 lemmas in the entire 1 million words corpus. Of the 4854 Chinese lemmas, only 62 of them have 3 or more framesets. In contrast, 294 lemmas have 3 or more framesets in the English Propbank.

[3]The small *pro* and big *PRO* distinction made in the Chinese Treebank is exploratory in nature. The idea is that it is easier to erase this distinction if it turns out to be implausible or infeasible than to add it if it turns out to be important.

Verb	English senses	Chinese translations
appear	be or have a quality of being	显得, 呈现
	come forth, become known or visible, physically or figuratively	出现,呈现
	present oneself formally, usually in a legal setting	露面
fight	combat or oppose	打好, 战斗, 抗
	strive, make a strenuous effort	奋斗
	promote, campaign or crusade	奋斗
join	connect, link or unite separate things, physically or abstractly	衔接, 接轨
	enlist or accept membership within some group or organization	走进, 参加, 加入
	participate with someone else in some event	同...一道, 同...一起
realize	be congnizant of, comprehend, perceive	认识,意识
	actualize, make real	实现
	take in , earn, acquire	实现
pass	tavel by	经
	clear, come through, succeed	通过
	elapse, happen	过去, 期满
	communicate	传出
settle	resolve, finalize, accept	解决
	reside, inhabit	进驻, 落户
raise	increase	提高
	lift, elevate, orient upwards	仰
	collect, levy	募集, 筹集, 筹措
	inovke, elicit, set off	提, 提出

Table 1: English verbs and their translations in the parallel Propbank

In our sense-tagging part of the project, we have been using manual groupings of the English WordNet senses. These groupings were previously shown to reconcile a substantial portion of the tagging disagreements, raising inter-annotator agreement from 71% in the case of fine-grained WordNet senses to 82% in the case of grouped senses for the Senseval 2 English data (Palmer et al., to appear), and currently to 89% for 93 new verbs (almost 12K instances) (Palmer et al., 2004). The question which arises, however, is how useful these grouped senses are and whether the level of granularity which they provide is sufficient for such applications as machine translation from English to Chinese.

In a preliminary investigation, we randomly selected 7 verbs and 5 nouns and looked at their corresponding translations in the Chinese Propbank. As the tables below show, for 6 verbs (join, pass, settle, raise, appear, fight) and 3 nouns (resolution, organization, development), grouped English senses map to unique Chinese translation sets. For a few examples, which include realize and party, grouped senses map to the same word in Chinese, preserving the ambiguity. This investigation justifies the appropriateness of the grouped sense tags, and indicates potential for providing a useful level of granularity for MT.

3.3 Discourse connectives

Another component of the Chinese / English Parallel Propbank II is the annotation of dis-course connectives for both Chinese corpus and its English translation. Like the other two components, the annotation is performed on the first 100K words of the Parallel Chinese English Treebank. The annotation of Chinese discourse connectives follows in large part the theoretic assumptions and annotation practices of the English Penn Discourse Project (PDTB) (Miltsakaki et al., 2004). Adaptations are made only when they are warranted by the linguistic facts of Chinese. While the English PTDB annotates both explicit and implicit discourse connectives, our ini-

Noun	English senses	Chinese translations
organization	individuals working together	组织,机构,单位
	event: putting things together	筹组
	state: the quality of being well-organization	组织
party	event: an occasion on which people can assemble for social interaction and entertainment	会
	political organization	党派
	a band of people associated temporarily in some activity	方
	person or side in legal context	
investment	time or money risked in hopes of profit	投资,资
	the act of investing	投资
development	the process of development	开发,发展
	the act of development	发展
resolution	a formal declaration	协议,决定
	coming to a solution	解决

Table 2: English nouns and their translations in the parallel Propbank

tial focus is on explicit discourse connectives. Explicit discourse connectives include subordinate (8) and coordinate conjunctions (9) as well as discourse adverbials (10). While subordinate and coordinate conjunctions are easy to understand, discourse adverbials need a little more elaboration. Discourse adverbials differ from other adverbials in that they relate two propositions. Typically one can be found in the immediate context while the other may need to be identified in the previous discourse.

(8) [*arg1* 台湾/Taiwan 商人/businessman] [**conn** 虽然/although] [*arg1* 生活/live 在/at 外/foreign land], [*arg2* 还是/still 很/very 注重/stress 孩子/child 教育/education].
"Although these Taiwan businessmen live away from home, they still stress the importance of their children's education."

(9) [*arg1* 东亚/East 各/every 国/country 间/among 并非/not really 完全/completely 没有/not have 矛盾/conflict 和/and 分歧/difference], [**conn** 但是/but] [*arg2* 为了/for 保障/protect 东亚/East Asia 各/every 国/country 的/DE 利益/interest, 必须/must 进一步/further 加强/strengthen 东亚/East Asia 合作/cooperation].
"It is not really true that there are no conflicts and differences among the East Asian countries, but in order to protect their common interest, they must cooperate."

(10) [*arg1* 浦东/Pudong 开发/development 是/BE 一/one 项/CL 振兴/invigorate 上海/Shanghai 的/DE 跨/across 世纪/century 工程/project], [**conn** 因此/therefore] [*arg2* 大量/large quantity 出现/appear 的/DE 是/BE 新/new 问题/problem]. "The development of Pudong, a project de-signed to invigorate Shanghai, spans over different centuries. Therefore, new problems occur in large quantities."

The annotation of the discourse connectives in a parallel English Chinese Propbank exposes interesting correspondences between English and Chinese discourse connectives. The examples in (11) show that "结果" is polysemous and corresponds with different expressions in English. It is a noun meaning "result" in (11a), where it is not a discourse connective. In (11b) it means "in the end", invoking a contrast between what has been planned and how the actual result turned out. In (11c) it means "as a result", expressing causality between the cause and the result.

(11) a. 实行/adopt "戒急用忍/go slow" 的/DE 政策/policy, 结果/**result** 是/BE 凭白/unnecessarily 丢失/lose 在/at 大陆/mainland 的/DE 商机/business opportunity.
"The result of adopting the 'go slow' policy is unnecessarily losing business opportunities in the mainland."

b. 纤维所/fiber institute 计划/plan 招收/enroll 十/10 名/CL 学生/student, 结果/**in the end** 只/only 有/have 二十/20 人/person 报名/register.
"The fiber institute planned to enroll 10 students. In the end, only 20 people registered to take the exam."

c. 学校/school 不/not 教/teach 理财/finance management, 一般/ordinary 人/people 又/and 有/have 这/this 方面/aspect 的/DE 需求/need, 结果/**as a result**, 报章/newspaper 上/on 各/every 种/kind 专栏/column 就/then 成为/become 资讯/information 的/DE 主要/main 来源/source.
"The school does not teach finance management and

ordinary people have this need. As a result, the different kinds of columns in the newspaper become the main source of information."

4 Conclusion

This paper presented preliminary results of the parallel PropBank II project. It highlighted some interesting aspects of the differences between English and Chinese, which play an important role for MT and other applications. Some of the questions addressed had to do with how events are lexicalized and aligned in the two languages, which level of sense granularity is needed for MT from English to Chinese, and highlighting notable differences between discourse connectives in the two languages. Further investigation and alignment of the parallel corpus, as well as richer annotation, will reveal other interesting phenomena.

References

Olga Babko-Malaya and Martha Palmer. 2005. Proposition Bank II: Delving Deeper. In *Frontiers in Corpus Annotation, Workshop in conjunction with HLT/NAACL 2004*, Boston, Massachusetts.

Olga Babko-Malaya, Martha Palmer, Nianwen Xue, Aravind Joshi, and Seth Kulick. 2004. Exploiting Interactions between Different Types of Semantic Annotation. In *Proceeding of ICWS-6*, Tilburg, The Netherlands.

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL*, Singapore.

E. Charniak. 2001. Immediate-head Parsing for Language Models. In *ACL-01*.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

M. Ellsworth, K. Erk, P. Kingsbury, and S. Pado. 2004. PropBank, SALSA and FrameNet: How design determines product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal.

Charles J. Fillmore and B. T. Atkins. 1998. FrameNet and lexical relevantce. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Eva Hajicova and Iyona Kucerova. 2002. Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 846–851.

Christopher R. Johnson, Charles J. Fillmore, Miriam R. L. Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J. Wood. 2002. FrameNet: Theory and Practice, Version 1.0, www.icsi.berkeley.edu/framenet.

M. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, et al. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proc of ARPA speech and Natural language workshop*.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Martha Palmer, Olga Babko-Malaya, and Hoa Dang. 2004. Different Sense Granularities for Different Applications. In *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems*, Boston, Mass.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. to appear. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. To appear. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*.

Nianwen Xue. To appear. Annotating the Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor, Michigan.

Semantically Rich Human-Aided Machine Annotation

Marjorie McShane, Sergei Nirenburg, Stephen Beale, Thomas O'Hara
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, Maryland, 21250 USA
{marge, sergei, sbeale, tomohara}@umbc.edu

Abstract

This paper describes a semantically rich, human-aided machine annotation system created within the Ontological Semantics (OntoSem) environment using the DEKADE toolset. In contrast to mainstream annotation efforts, this method of annotation provides more information at a lower cost and, for the most part, shifts the maintenance of consistency to the system itself. In addition, each tagging effort not only produces knowledge resources for that corpus, but also leads to improvements in the knowledge environment that will better support subsequent tagging efforts.

1 Introduction

Corpus tagging is a prerequisite for many machine learning methods in NLP but has the drawbacks of high cost, inter-annotator inconsistency and the insufficient treatment of meaning. A tagging approach that strives to ameliorate all of these drawbacks is semantically rich, human-aided machine annotation (HAMA), implemented in the OntoSem (Ontological Semantics) environment using a toolset called DEKADE: the Development, Evaluation, Knowledge Acquisition and Demonstration Environment of OntoSem.

In brief, the OntoSem text analyzer takes as input open text and outputs a text-meaning representation (TMR) that represents its meaning using an ontologically grounded, language-independent metalanguage (see Nirenburg and Raskin 2004). Since the processing leading up to the production of TMR includes, in addition to semantic analysis proper, preprocessing (roughly, segmentation, treatment of named entities and morphology) and syntactic analysis, the overall annotation of text in this approach includes tags relating to all of the above levels. Since the typical input for analysis in our practice is genuine sentences, which are on average 25 words long and contain all manner of complex phenomena, it is not uncommon for the automatically generated TMRs to contain errors. These errors—which can occur at the level of preprocessing, syntactic analysis or semantic analysis—can be corrected manually using the DEKADE environment, yielding "gold standard" output. Making a human the final arbiter in the process means that such long-term complexities as treatment of metaphor, metonymy, PP-attachment, difficult cases of reference resolution and others can be resolved locally while we work on fundamental, implementable automatic solutions.

In this paper we describe the OntoSem/DEKADE environment for the creation of gold standard TMRs, which supports the first ever annotation effort that:

- produces structures that can be used as input for both text generators and general reasoning systems: semantically rich representations of the meaning of text written in a language-independent metalanguage; these representations cover entities, propositions, relations, attributes, speaker attitudes, modalities, polarity, discourse relations, time, reference relations, and more;
- produces semantic tagging of text largely automatically, thus making more realistic and affordable the tagging of large amounts of text in finite time;
- almost fully circumvents the pitfalls of manual tagging, including human tagger errors and inconsistencies;
- produces richer semantic annotations than manual tagging realistically could, since manipulating large and complex static knowl-

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 68–75,
Ann Arbor, June 2005.

edge sources would be impossible for humans if starting from scratch (i.e., our methodology effectively turns an essay question into a multiple choice one, with most of the correct answers already provided);

- incorporates humans as final arbiters for output of three stages of text analysis (preprocessing, syntactic analysis and semantic analysis), thus maximally leveraging the automated capacity of the system but not requiring of it blanket coverage at this point in its development;
- promises to reduce, over time, the dependence on human input because an important side effect of the operation of the human-assisted machine annotation approach is enhancement of the static knowledge resources – the lexicon and the ontology – underlying the OntoSem analyzer, so that the quality of automatic text analysis will grow as the HAMA system operates, leading to an ever improving quality of raw, unedited TMRs;
- (as a corollary to the previous point) becomes more cost-efficient over time; and
- can be cost-effectively extended to other languages (including less commonly taught languages), with much less work than was required for the first language since many of the necessary resources are language-independent.

Our approach to text analysis is a hybrid of knowledge-based and corpus-based, stochastic methods.

In the remainder of the paper we will briefly describe the lay of the land in text annotation (Section 2), the OntoSem environment (Section 3), the DEKADE environment for creating gold-standard TMRs from automatically generated ones (Section 4), the portability of OntoSem to other languages (Section 5), and the broader implications of this R&D effort (Section 6).

2 The Lay of the Land in Annotation

In addition to the well-known bottlenecks of cost and inconsistency, it is widely assumed that low-level (only syntactic or "light semantic") tagging is either sufficient or inevitable due to the complexity of semantic tagging. Past and ongoing tagging efforts share this point of departure.

Numerous projects have striven to achieve text annotation via a simpler task, like translation, sometimes assuming that one language has already been tagged (e.g., Pianta and Bentivogli 2003, and references therein). But results of such efforts are either of low quality, light semantic depth, or remain to be reported. Of significant interest is the porting of annotations across languages: for example, Yarowsky et al. 2001 present a method for automatic tagging of English and the projection of the tags to other languages; however, these tags do not include semantics.

Post-editing of automatic annotation has been pursued in various projects (e.g., Brants 2000, and Marcus et al. 1993). The latter group did an experiment early on in which they found that "manual tagging took about twice as long as correcting [automated tagging], with about twice the inter-annotator disagreement rate and an error rate that was about 50% higher" (Marcus et al. 1993). This conclusion supports the pursuit of automated tagging methods. The difference between our work and the work in the above projects, however, is that syntax for us is only a step in the progression toward semantics.

Interesting time- and cost-related observations are provided in Brants 2000 with respect to the manual correction of automated POS and syntactic tagging of a German corpus (semantics is not addressed). Although these tasks took approximately 50 seconds per sentence, with sentences averaging 17.5 tokens, the actual cost in time and money puts each sentence at 10 minutes, by the time two taggers carry out the task, their results are compared, difficult issues are resolved, and taggers are trained in the first place. Notably, however, this effort used students as taggers, not professionals. We, by contrast, use professionals to check and correct TMRs and thus reduce to practically zero the training time, the need for multiple annotators (provided the size of a typical annotation task is commensurate with those in current projects), and costly correction of errors.

Among past projects that have addressed semantic annotation are the following:

1. Gildea and Jurafsky (2002) created a stochastic system that labels case roles of predicates with either abstract (e.g., AGENT, THEME) or domain-specific (e.g., MESSAGE, TOPIC) roles. The system

trained on 50,000 words of hand-annotated text (produced by the FrameNet project). When tasked to segment constituents and identify their semantic roles (with fillers being undisambiguated textual strings) the system scored in the 60's in precision and recall. Limitations of the system include its reliance on hand-annotated data, and its reliance on prior knowledge of the predicate frame type (i.e., it lacks the capacity to disambiguate productively). Semantics in this project is limited to case-roles.

2. The goal of the "Interlingual Annotation of Multilingual Text Corpora" project (http://aitc.aitcnet.org/nsf/iamtc/) is to create a syntactic and semantic annotation representation methodology and test it out on six languages (English, Spanish, French, Arabic, Japanese, Korean, and Hindi). The semantic representation, however, is restricted to those aspects of syntax and semantics that developers believe can be consistently handled well by hand annotators for many languages. The current stage of development includes only syntax and light semantics – essentially, thematic roles.

3. In the ACE project (http://www.ldc.upenn.edu/Projects/ACE/intro.html), annotators carry out manual semantic annotation of texts in English, Chinese and Arabic to create training and test data for research task evaluations. The downside of this effort is that the inventory of semantic entities, relations and events is very small and therefore the resulting semantic representations are coarse-grained: e.g., there are only five event types. The project description promises more fine-grained descriptors and relations among events in the future.

4. Another response to the insufficiency of syntax-only tagging is offered by the developers of PropBank, the Penn Treebank semantic extension. Kingsbury et al. 2002 report: "It was agreed that the highest priority, and the most feasible type of semantic annotation, is coreference and predicate argument structure for verbs, participial modifiers and nominalizations", and this is what is included in PropBank.

To summarize, previous tagging efforts that have addressed semantics at all have covered only a relatively small subset of semantic phenomena. OntoSem, by contrast, produces a far richer annotation, carried out largely automatically, within an environment that will improve over time and with use.

3 A Snapshot of OntoSem

OntoSem is a text-processing environment that takes as input unrestricted raw text and carries out preprocessing, morphological analysis, syntactic analysis, and semantic analysis, with the results of semantic analysis represented as formal text-meaning representations (TMRs) that can then be used as the basis for many applications (for details, see, e.g., Nirenburg and Raskin 2004, Beale et al. 2003). Text analysis relies on:

- The OntoSem language-independent ontology, which is written using a metalanguage of description and currently contains around 6,000 concepts, each of which is described by an average of 16 properties.
- An OntoSem lexicon for each language processed, which contains syntactic and semantic zones (linked using variables) as well as calls for procedural semantic routines when necessary. The semantic zone most frequently refers to ontological concepts, either directly or with property-based modifications, but can also describe word meaning extra-ontologically, for example, in terms of modality, aspect, time, etc. The current English lexicon contains approximately 25,000 senses, including most closed-class items and many of the most frequent and polysemous verbs, as targeted by corpus analysis. (An extensive description of the lexicon, formatted as a tutorial, can be found at http://ilit.umbc.edu.)
- An onomasticon, or lexicon of proper names, which contains approximately 350,000 entries.
- A fact repository, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3366 is the 3366th instantiation of the concept SPEECH-ACT in the world model constructed during the processing of some given text(s)).
- The OntoSem syntactic-semantic analyzer, which covers preprocessing, syntactic analysis, semantic analysis, and the creation of TMRs. Instead of using a large, monolithic grammar of a language, which leads to ambiguity and inefficiency, we use a special lexicalized grammar created on the fly for each input sentence (Beale, et. al. 2003). Syntactic rules are generated from the lexicon entries of each of the words in the sentence, and are supplemented by a small inventory of generalized rules. We augment this

basic grammar with transformations triggered by words or features present in the input sentence.
- The TMR language, which is the metalanguage for representing text meaning.

Creating gold standard TMRs involves running text through the OntoSem processors and checking/correcting the output after three stages of analysis: preprocessing, syntactic analysis, and semantic analysis. These outputs can be viewed and edited as text or as visual representations through the DEKADE interface. Although the gold standard TMR itself does not reflect the results of preprocessing or syntactic analysis, the gold standard results of those stages of processing are stored in the system and can be converted into a more traditional annotation format.

4 TMRs in DEKADE

TMRs represent propositions connected by discourse relations (since space permits only the briefest of descriptions, interested readers are directed to Nirenburg and Raskin 2004, Chapter 6 for details). Propositions are headed by instances of ontological concepts, parameterized for modality, aspect, proposition time, overall TMR time, and style. Each proposition is related to other instantiated concepts using ontologically defined relations (which include case roles and many others) and attributes. Coreference links form an additional layer of linking between instantiated concepts. OntoSem microtheories devoted to modality, aspect, time, style, reference, etc., undergo iterative extensions and improvements in response to system needs as diagnosed during the processing of actual texts.

We use the following sentence to walk through the processes of automatically generating TMRs and viewing/editing those TMRs to create a gold-standard annotated corpus.

The Iraqi government has agreed to let U.S. Representative Tony Hall visit the country to assess the humanitarian crisis.

Preprocessor. The preprocessor identifies the root word, part of speech and morphological features of each word; recognizes sentence boundaries, named entities, dates, times and numbers; and for named entities, determines the ontological type (i.e. HUMAN, PLACE, ORGANIZATION, etc.) of the entity as well as its subparts (e.g., the first, last, and middle names of a person). For the semi-automatic creation of gold standard TMRs, much ambiguity can be removed at small cost by allowing people to correct spurious part-of-speech tags, number and date boundaries, etc., through the DEKADE environment at the preprocessor stage (see Figure 1). Clicking on w+ permits a new POS tag/analysis, and clicking on w-, the more common action, removes spurious analyses. Preprocessor correction is a conceptually simple and logistically fast task that can be carried out by less trained, and therefore less expensive, annotators.

+	-	Root	Pos	Features
w+	w-	THE	ART	NIL
w+	w-	IRAQI	N	((TYPE COUNTRY) (TYPE PN))
w+	w-	GOVERNM...	N	NIL
w+	w-	HAVE	V	((PERSON THIRD) (TENSE PRESENT))
w+	w-	AGREE	V	((FORM PAST-PARTICIPLE) (TENSE PAST))
w+	w-	TO	INF	NIL
w+	w-	TO	PREP	NIL
w+	w-	LET	V	((FORM INFINITIVE) (TENSE PRESENT))
w+	w-	LET	N	NIL
w+	w-	U*PERIOD...	N	((TYPE PN))
w+	w-	REPRESE...	N	NIL
w+	w-	REPRESE...	ADJ	NIL
w+	w-	TONY-HALL	N	((TYPE NAME) (TYPE PN))
w+	w-	*PERSON*	N	((TYPE PN) (TYPE NAME) (LAST HALL) (FIRST TONY))
w+	w-	VISIT	V	((FORM INFINITIVE) (TENSE PRESENT))
w+	w-	VISIT	N	NIL
w+	w-	THE	ART	NIL

Save

Figure 1. Preprocesor Output Editor.

Syntax. Syntax output can be viewed and edited in text or graphic form. The graphic viewer/editor presents the sentence using the traditional metaphor of color-coded labeled arcs. Mouse clicks show the components of arcs, permit arcs to be deleted along with the orphans they would leave, allow for the edges of arcs to be moved, etc. (no graphic of the syntax or semantics browsers/editors are provided due to space constraints).

One common error in syntax output is spurious parses due to contextually incorrect POS or feature analysis. As shown above, this can be fixed from the outset by correcting the preprocessor. However, since the preprocessor will always contain spurious analyses that can usually be removed automatically by the syntactic analyzer, it is not necessarily most time efficient to always start with preprocessor editing. A more difficult, long-term research issue is genuine ambiguity caused, for example, by PP-attachments. While such issues are

not likely to be solved computationally in the short term, they can be easily resolved when humans are used as the final arbiters in the creation of gold standard TMRs.

When the correct parse is not included in the syntactic output, either the necessary lexical knowledge is lacking (i.e. there is an unknown word or word sense), or an unknown grammatical construction has been used. While the syntax-editing interface permits spot-correction of the problem by the addition of the necessary arc(s), a more fundamental knowledge-building approach is generally preferred – except when the input is non-standard, in which case systemic modifications are avoided.

Semantics. Within the OntoSem environment, there are two stages of text-meaning representations (TMRs): basic and extended. The basic TMR shows the basic ontological mappings and dependency structure, whereas the extended TMR shows the results of procedural semantics, including reference resolution, reasoning about time relations, etc. The basic and extended stages of TMR creation can be viewed and edited separately within DEKADE.

TMRs can be viewed and edited in text format or graphically. In the latter, concepts are shown as nodes and properties are shown as lines connecting them. A pretty-printed view of the textual extended TMR for our sample sentence, repeated for convenience, is as follows (concept names are in small caps; instance numbers are appended to them).

The Iraqi government has agreed to let U.S. Representative Tony Hall visit the country to assess the humanitarian crisis.

AGREE-268
  textpointer agree
  THEME MODALITY-200
  AGENT GOVERNMENTAL-ORGANIZATION-41
  TIME (< FIND-ANCHOR-TIME)
GOVERNMENTAL-ORGANIZATION-41
  textpointer government
  RELATION NATION-56
  AGENT-OF AGREE-268
NATION-56
  textpointer Iraq
  RELATION GOVERNMENTAL-ORGANIZATION-41
MODALITY-200
  textpointer let
  TYPE permissive
  SCOPE TRAVEL-EVENT-272
  VALUE 1
TRAVEL-EVENT-272
  textpointer visit
  AGENT SENATOR-447[1]
  DESTINATION NATION-57
  PURPOSE EVALUATE-69
  SCOPE-OF MODALITY-200
SENATOR-447
  textpointer Representative Tony Hall[2]
  REPRESENTATIVE-OF NATION-40
NATION-40
  textpointer U.S.
  REPRESENTED-BY SENATOR-447
NATION-57
  textpointer country
  COREFER NATION-56
EVALUATE-69
  AGENT SENATOR-447
  THEME DISASTER-EVENT-2
DISASTER-EVENT-2
  BENEFICIARY SET-23
  THEME-OF EVALUATE-69
SET-23
  MEMBER-TYPE HUMAN-1342
  BENEFICIARY-OF DISASTER-EVENT-2

Within the graphical browser, clicking on concept names or properties permits them to be deleted, edited, or permits new ones to be added. It also shows the expansion of any concept in text format.

Evaluating and editing the semantic output is the most challenging aspect of creating gold standard TMRs, since creating formal semantic representations is arguably one of the most difficult tasks in all of NLP. If a knowledge engineer determines that some aspect of the semantic representation is incorrect, the problems can be corrected locally or by editing the knowledge resources and rerunning the analyzer. Local corrections are used, for example, in cases of metaphor and metonymy, which we do not record in our knowledge resources (we are working on a microtheory of tropes but it is not yet implemented). In all other cases, resource supplementation is preferred; it can be carried out either immediately or the problem can be fixed locally, in which case a request will be sent to a knowledge acquirer to carry out the necessary resource enhancements.

[1] The concept SENATOR is defined as a member of a legislative assembly.
[2] Collocations of SOCIAL-ROLE + personal name are handled by the preprocessor.

Striking the balance between short-term goals (a gold standard TMR for the given text) and long-term goals (better analysis of any text in the future) is always a challenge. For example, if a text contained the word *grass* in the sense of 'marijuana', and if the lexicon lacked the word 'grass' altogether, we would want to acquire the meaning 'green lawn cover' as well; however, doing this without constraint could mean getting bogged down by knowledge acquisition (as with the dozens of idiomatic uses of 'have') at the expense of actually producing gold-standard TMRs. There are also cases in which a local solution to semantic representation is very easy whereas a fundamental, machine-reproducible solution is very difficult. Consider the case of relative expressions, like *respective* and *respectively*, as used in *Smith and Matthews pleaded innocent and guilty,* ***respectively***. Manually editing a TMR such that the appropriate properties are linked to their heads is quite simple, whereas writing a program for this non-trivial case of reference resolution is not. Thus, in some cases we push through gold standard TMR production while keeping track of – and developing as time permits – the more difficult aspects of text processing that will enhance TMR output in the future.

The gold standard TMR for the sentence discussed at length here was produced with only a few manual corrections: changing two part of speech tags and selecting the correct sense for one word. Work took less than the 10 minutes reported by Brants 2000 for their non-semantic tagging.

5 Porting to Other Languages

Recently the need for tagged corpora for less commonly taught languages has received much attention. While our group is not currently pursuing such languages, it has in the past: TMRs have been automatically generated for languages such as Chinese, Georgian, Arabic and Persian. We take a short tangent to explain how OntoSem/DEKADE can be extended, at relatively low cost, to the annotation of other languages – showing yet another way in which this approach to annotation reaches beyond the results for any given text or corpus.

Whereas it is typical to assume that lexicons are language-specific whereas ontologies are language-independent, most aspects of the semantic structures (*sem-strucs*) of OntoSem lexicon entries are actually language-independent, apart from the linking of specific variables to their counterparts in the syntactic structure. Stated differently, if we consider *sem-strucs* – no matter what lexicon they originate from – to be building blocks of the representation of *word meaning* (as opposed to concept meaning, as is done in the ontology), then we understand why building a large OntoSem lexicon for English holds excellent promise for future porting to other languages: most of the work is already done. This conception of cross-linguistic lexicon development derives in large part from the Principle of Practical Effability (Nirenburg and Raskin 2004), which states that what can be expressed in one language can *somehow* be expressed in all other languages, be it by a word, a phrase, etc. (Of course, it is not necessary that every nuanced meaning be represented in the lexicon of every language and, as such, there will be some differences in the lexical stock of each language: e.g., whereas German has a word for *white horse* which will be listed in its lexicon, English will not have such a lexical entry, the collocation *white horse* being treated compositionally.) We do not intend to trivialize the fact that creating a new lexicon is a lot of work. It is, however, compelling to consider that a new lexicon of the same quality of our OntoSem English one could be created with little more work than would be required to build a typical translation dictionary. In fact, we recently carried out an experiment on porting the English lexicon to Polish and found that a) much of it could be done semi-automatically and b) the manual work for a second language is considerably less than for the first language (for further discussion, see McShane et al. 2004).

To sum up, the OntoSem ontology and the DEKADE environment are equally suited to any language, and the OntoSem English lexicon and analyzer can be configured to new languages with much less work required than for their initial development. In short, semantic-rich tagging through TMR creation could be a realistic option for languages other than English.

6 Discussion

Lack of interannotator agreement presents a significant problem in annotation efforts (see, e.g., Marcus et al. 1993). With the OntoSem semi-automated approach, there is far less possibility of

interannotator disagreement since people only correct the output of the analyzer, which is responsible for consistent and correct deployment of the large and complex static resources: if the knowledge bases are held constant, the analyzer will produce the same output every time, ensuring reproducibility of the annotation.

Evaluation of annotation has largely centered upon the demonstration of interannotator agreement, which is at best a partial standard for evaluation. On the one hand, agreement among annotators does not imply the correctness of the annotations: all annotators could be mistaken, particularly as students are most typically recruited for the job. On the other hand, there are cases of genuine ambiguity, in which more than one annotation is equally correct. Such ambiguity is particularly common with certain classes of referring expressions, like *this* and *that*, which can refer to chunks of text ranging from a noun phrase to many paragraphs. Genuine ambiguity in the context of corpus tagging has been investigated by Poesio and Artstein (ms.), among others, who conclude, reasonably, that a system of tags must permit multiple possible correct coreference relations and that it is useful to evaluate coreference based on coreference chains rather than individual entities.

The abovementioned evidence suggests the need for ever more complex evaluation metrics which are costly to develop and deploy. In fact, evaluation of a complex tagging effort will be almost as complex as the core work itself. In our case, TMRs need to be evaluated not only for their correctness with respect to a given state of knowledge resources but also in the abstract. Speed of gold standard TMR creation must also be evaluated, as well as the number of mistakes at each stage of analysis, and the effect that the correction of output at one stage has on the next stage. No methods or standards for such evaluation are readily available since no work of this type has ever been carried out.

In the face of the usual pressures of time and manpower, we have made the programmatic decision not to focus on all types of evaluation but, rather, to concentrate our evaluation metrics on the correctness of the automated output of the system, the extent to which manual correction is needed, and the depth and robustness of our knowledge resources (see Nirenburg et al. 2004 for our first evaluation effort). We do not deny the ultimate desirability of additional aspects of evaluation in the future.

The main source of variation among knowledge engineers within our approach lies not in reviewing/editing annotations as such, but in building the knowledge sources that give rise to them. To take an actual example we encountered: one member of our group described the phrase *weapon of mass destruction* in the lexicon as BIOLOGICAL-WEAPON or CHEMICAL-WEAPON, while another described it as a WEAPON with the potential to kill a very large number of people/animals. While both of these are correct, they focus on different salient aspects of the collocation. Another example of potential differences at the knowledge level has to do with grain size: whereas one knowledge engineer reviewing a TMR might consider the current lexical mapping of *neurosurgeon* to SURGEON perfectly acceptable, another might consider that this grain size is too rough and that, instead, we need a new concept NEUROSURGEON, whose special properties are ontologically defined. Such cases are to be expected especially as we work on new specialized domains which put greater demands on the depth of knowledge encoded about relevant concepts.

There has been some concern that manual editing of automated annotation can introduce bias. Unfortunately, completely circumventing bias in semantic annotation is and will remain impossible since the process involves semantic interpretation, which often differs among individuals from the outset. As such, even agreements among annotators can be questioned by a third (fourth, etc.) party.

At the present stage of development, the TMR together with the static (ontology, lexicons) and dynamic (analyzer) knowledge sources that are used in generating and manipulating it, already provide substantial coverage for a broad variety of semantic phenomena and represent in a compact way practically attainable solutions for most issues that have concerned the computational linguistics and NLP community for over fifty years. Our TMRs have been used as the substrate for question-answering, MT, knowledge extraction, and were also used as the basis for reasoning in the question-answering system AQUA, where they supplied knowledge to enable the operation of the JTP (Fikes et al., 2003) reasoning module.

We are creating a database of TMRs paired with their corresponding sentences that we believe

will be a boon to machine learning research. Repeatedly within the ML community, the creation of a high quality dataset (or datasets) for a particular domain has sparked development of applications, such as learning semantic parsers, learning lexical items, learning about the structure of the underlying domain of discourse, and so on. Moreover, as the quality of the raw TMRs increases due to general improvements to the static resources (in part, as side effects of the operation of the HAMA process) and processors (a long-term goal), the net benefit of this approach will only increase, as the production rate of gold-standard TMRs will increase thus lowering the costs.

TMRs are a useful medium for semantic representation in part because they can capture any content in any language, and even content not expressed in natural language. They can, for example, be used for recording the interim and final results of reasoning by intelligent agents. We fully expect that, as the actual coverage in the ontology and the lexicons and the quality of semantic analysis grows, the TMR format will be extended to accommodate these improvements. Such an extension, we believe, will largely involve movement toward a finer grain size of semantic description, which the existing formalism should readily allow. The metalanguage of TMRs is quite transparent, so that the task of converting them into a different representation language (e.g., OWL) should not be daunting.

References

Stephen Beale, Sergei Nirenburg and Marjorie McShane. 2003. Just-in-time grammar. *Proceedings of the 2003 International Multiconference in Computer Science and Computer Engineering*. Las Vegas, Nevada.

Thorsten Brants. 2000. Inter-annotator agreement for a German newspaper corpus. *LREC-2000.* Athens, Greece.

Richard Fikes, Jessica Jenkins and Gleb Frank. 2003. JTP: A system architecture and component library for hybrid reasoning. *Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics*. Orlando, Florida, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28:3, 245-288.

Paul Kingsbury, Martha Palmer and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. (http://www.cis.upenn.edu/~ace/ HLT2002-propbank.pdf.)

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19.

Marjorie McShane, Margalit Zabludowski, Sergei Nirenburg and Stephen Beale. 2004. OntoSem and SIMPLE: Two multi-lingual world views. *Proceedings of ACL-2004 Workshop on Text Meaning and Interpretation*. Barcelona, Spain.

Sergei Nirenburg, Stephen Beale and Marjorie McShane. 2004. Evaluating the performance of the OntoSem semantic analyzer. *Proceedings of the ACL Workshop on Text Meaning Representation*. Barcelona, Spain.

Sergei Nirenburg and Victor Raskin. 2004. *Ontological Semantics*. The MIT Press.

Emanuele Pianta and Luisa Bentivogli. 2003. Translation as annotation. *Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy."* Pisa, Italy.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. *Proceedings of the ACL 2005 Workshop "Frontiers in Corpus Annotation II, Pie in the Sky"*.

David Yarowsky, Grace Ngai and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. *Proceedings of HLT 2001, First International Conference on Human Language Technology Research,* San Diego, California, USA.

The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account

Massimo Poesio and Ron Artstein
University of Essex,
Language and Computation Group / Department of Computer Science
United Kingdom

Abstract

We report the results of a study of the reliability of anaphoric annotation which (i) involved a substantial number of naive subjects, (ii) used Krippendorff's α instead of K to measure agreement, as recently proposed by Passonneau, and (iii) allowed annotators to mark anaphoric expressions as ambiguous.

1 INTRODUCTION

We tackle three limitations with the current state of the art in the annotation of anaphoric relations. The first problem is the lack of a truly systematic study of agreement on anaphoric annotation in the literature: none of the studies we are aware of (Hirschman, 1998; Poesio and Vieira, 1998; Byron, 2003; Poesio, 2004) is completely satisfactory, either because only a small number of coders was involved, or because agreement beyond chance couldn't be assessed for lack of an appropriate statistic, a situation recently corrected by Passonneau (2004). The second limitation, which is particularly serious when working on dialogue, is our still limited understanding of the degree of agreement on references to abstract objects, as in discourse deixis (Webber, 1991; Eckert and Strube, 2001).

The third shortcoming is a problem that affects all types of semantic annotation. In all annotation studies we are aware of,[1] the fact that an expression may not have a unique interpretation in the context of its occurrence is viewed as a problem with the annotation scheme, to be fixed by, e.g., developing suitably underspecified representations, as done particularly in work on wordsense annotation (Buitelaar, 1998; Palmer et al., 2005), but also on dialogue act tagging. Unfortunately, the underspecification solution only genuinely applies to cases of polysemy, not homonymy (Poesio, 1996), and anaphoric ambiguity is not a case of polysemy. Consider the dialogue excerpt in (1):[2] it's not clear to us (nor was to our annotators, as we'll see below) whether the demonstrative *that* in utterance unit 18.1 refers to the 'bad wheel' or 'the boxcar'; as a result, annotators' judgments may disagree – but this doesn't mean that the annotation scheme is faulty; only that what is being said is genuinely ambiguous.

(1)
```
18.1 S: ....
18.6    it turns out that the boxcar
        at Elmira
18.7    has a bad wheel
18.8    and they're .. gonna start
        fixing that at midnight
18.9    but it won't be ready until 8
19.1 M: oh what a pain in the butt
```

This problem is encountered with all types of annotation; the view that all types of disagreement indicate a problem with the annotation scheme–i.e., that somehow the problem would disappear if only we could find the right annotation scheme, or concentrate on the 'right' types of linguistic judgments– is, in our opinion, misguided. A better approach

[1] The one exception is Rosenberg and Binkowski (2004).

[2] This example, like most of those in the rest of the paper, is taken from the first edition of the TRAINS corpus collected at the University of Rochester (Gross et al., 1993). The dialogues are available at `ftp://ftp.cs.rochester.edu/pub/papers/ai/92.tn1.trains_91_dialogues.txt`.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 76–83,
Ann Arbor, June 2005.

is to find when annotators disagree because of intrinsic problems with the text, or, even better, to develop methods to identify genuinely ambiguous expressions–the ultimate goal of this work.

The paper is organized as follows. We first briefly review previous work on anaphoric annotation and on reliability indices. We then discuss our experiment with anaphoric annotation, and its results. Finally, we discuss the implications of this work.

2 ANNOTATING ANAPHORA

It is not our goal at this stage to propose a new scheme for annotating anaphora. For this study we simply developed a coding manual for the purposes of our experiment, broadly based on the approach adopted in MATE (Poesio et al., 1999) and GNOME (Poesio, 2004), but introducing new types of annotation (ambiguous anaphora, and a simple form of discourse deixis) while simplifying other aspects (e.g., by not annotating bridging references).

The task of 'anaphoric annotation' discussed here is related, although different from, the task of annotating 'coreference' in the sense of the so-called MUCSS scheme for the MUC-7 initiative (Hirschman, 1998). This scheme, while often criticized, is still widely used, and has been the basis of coreference annotation for the ACE initiative in the past two years. It suffers however from a number of problems (van Deemter and Kibble, 2000), chief among which is the fact that the one semantic relation expressed by the scheme, ident, conflates a number of relations that semanticists view as distinct: besides COREFERENCE proper, there are IDENTITY ANAPHORA, BOUND ANAPHORA, and even PREDICATION. (Space prevents a fuller discussion and exemplification of these relations here.)

The goal of the MATE and GNOME schemes (as well of other schemes developed by Passonneau (1997), and Byron (2003)) was to devise instructions appropriate for the creation of resources suitable for the theoretical study of anaphora from a linguistic / psychological perspective, and, from a computational perspective, for the evaluation of anaphora resolution and referring expressions generation. The goal is to annotate the *discourse model* resulting from the interpretation of a text, in the sense both of (Webber, 1979) and of dynamic theories of anaphora (Kamp and Reyle, 1993). In order to do this, annotators must first of all identify the noun phrases which either introduce new discourse entities (discourse-new (Prince, 1992)) or are mentions of previously introduced ones (discourse-old), ignoring those that are used predicatively. Secondly, annotators have to specify which discourse entities have the same interpretation. Given that the characterization of such discourse models is usually considered part of the area of the semantics of anaphora, and that the relations to be annotated include relations other than Sidner's (1979) COSPECIFICATION, we will use the term ANNOTATION OF ANAPHORA for this task (Poesio, 2004), but the reader should keep in mind that we are not concerned only with nominal expressions which are lexically anaphoric.

3 MEASURING AGREEMENT ON ANAPHORIC ANNOTATION

The agreement coefficient which is most widely used in NLP is the one called K by Siegel and Castellan (1988). Howewer, most authors who attempted anaphora annotation pointed out that K is not appropriate for anaphoric annotation. The only sensible choice of 'label' in the case of (identity) anaphora are anaphoric chains (Passonneau, 2004); but except when a text is very short, few annotators will catch all mentions of the same discourse entity–most forget to mark a few, which means that agreement as measured with K is always very low. Following Passonneau (2004), we used the coefficient α of Krippendorff (1980) for this purpose, which allows for partial agreement among anaphoric chains.[3]

3.1 Krippendorf's alpha

The α coefficient measures agreement among a set of coders *C* who assign each of a set of items *I* to one of a set of distinct and mutually exclusive categories *K*; for anaphora annotation the coders are the annotators, the items are the markables in the text, and the categories are the emerging anaphoric chains. The coefficient measures the observed disagreement between the coders D_o, and corrects for

[3] We also tried a few variants of α, but these differed from α only in the third to fifth significant digit, well below any of the other variables that affected agreement. In the interest of space we only report here the results obtained with α.

chance by removing the amount of disagreement expected by chance D_e. The result is subtracted from 1 to yield a final value of agreement.

$$\alpha = 1 - \frac{D_o}{D_e}$$

As in the case of K, the higher the value of α, the more agreement there is between the annotators. $\alpha = 1$ means that agreement is complete, and $\alpha = 0$ means that agreement is at chance level.

What makes α particularly appropriate for anaphora annotation is that the categories are not required to be disjoint; instead, they must be ordered according to a DISTANCE METRIC–a function **d** from category pairs to real numbers that specifies the amount of dissimilarity between the categories. The distance between a category and itself is always zero, and the less similar two categories are, the larger the distance between them. Table 1 gives the formulas for calculating the observed and expected disagreement for α. The amount of disagreement for each item $i \in I$ is the arithmetic mean of the distances between the pairs of judgments pertaining to it, and the observed agreement is the mean of all the item disagreements. The expected disagreement is the mean of the distances between all the judgment pairs in the data, without regard to items.

$$D_o = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_{i \in I} \sum_{k \in K} \sum_{k' \in K} \mathbf{n}_{ik} \mathbf{n}_{ik'} \mathbf{d}_{kk'}$$

$$D_e = \frac{1}{\mathbf{ic}(\mathbf{ic}-1)} \sum_{k \in K} \sum_{k' \in K} \mathbf{n}_k \mathbf{n}_{k'} \mathbf{d}_{kk'}$$

c	number of coders
i	number of items
$\mathbf{n}_{ik}$	number of times item i is classified in category k
$\mathbf{n}_k$	number of times any item is classified in category k
$\mathbf{d}_{kk'}$	distance between categories k and k'

Table 1: Observed and expected disagreement for α

3.2 Distance measures

The distance metric is not part of the general definition of α, because different metrics are appropriate for different types of categories. For anaphora annotation, the categories are the ANAPHORIC CHAINS: the sets of markables which are mentions of the same discourse entity. Passonneau (2004) proposes a distance metric between anaphoric chains based on the following rationale: two sets are minimally distant when they are identical and maximally distant when they are disjoint; between these extremes, sets that stand in a subset relation are closer (less distant) than ones that merely intersect. This leads to the following distance metric between two sets A and B.

$$\mathbf{d}_{AB} = \begin{cases} 0 & \text{if } A = B \\ 1/3 & \text{if } A \subset B \text{ or } B \subset A \\ 2/3 & \text{if } A \cap B \neq \emptyset, \text{ but } A \not\subset B \text{ and } B \not\subset A \\ 1 & \text{if } A \cap B = \emptyset \end{cases}$$

We also tested distance metrics commonly used in Information Retrieval that take the size of the anaphoric chain into account, such as Jaccard and Dice (Manning and Schuetze, 1999), the rationale being that the larger the overlap between two anaphoric chains, the better the agreement. Jaccard and Dice's set comparison metrics were subtracted from 1 in order to get measures of distance that range between zero (minimal distance, identity) and one (maximal distance, disjointness).

$$\mathbf{d}_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad \text{(Jaccard)}$$

$$\mathbf{d}_{AB} = 1 - \frac{2|A \cap B|}{|A| + |B|} \qquad \text{(Dice)}$$

The Dice measure always gives a smaller distance than the Jaccard measure, hence Dice always yields a higher agreement coefficient than Jaccard when the other conditions remain constant. The difference between Dice and Jaccard grows with the size of the compared sets. Obviously, the Passonneau measure is not sensitive to the size of these sets.

3.3 Computing the anaphoric chains

Another factor that affects the value of the agreement coefficient–in fact, arguably the most important factor–is the method used for constructing from the raw annotation data the 'labels' used for agreement computation, i.e., the anaphoric chains. We experimented with a number of methods. However, since the raw data are highly dependent on the annotation scheme, we will postpone discussing our chain construction methods until after we have described our experimental setup and annotation scheme. We will also discuss there how comparisons are made when an ambiguity is marked.

4 THE ANNOTATION STUDY

4.1 The Experimental Setup

Materials. The text annotated in the experiment was dialogue 3.2 from the TRAINS 91 corpus. Subjects were trained on dialogue 3.1.

Tools. The subjects performed their annotations on Viglen Genie workstations with LG Flatron monitors running Windows XP, using the MMAX 2 annotation tool (Müller and Strube, 2003).[4]

Subjects. Eighteen paid subjects participated in the experiment, all students at the University of Essex, mostly undergraduates from the Departments of Psychology and Language and Linguistics.

Procedure. The subjects performed the experiment together in one lab, each working on a separate computer. The experiment was run in two sessions, each consisting of two hour-long parts separated by a 30 minute break. The first part of the first session was devoted to training: subjects were given the annotation manual and taught how to use the software, and then annotated the training text together. After the break, the subjects annotated the first half of the dialogue (up to utterance 19.6). The second session took place five days later. In the first part we quickly pointed out some problems in the first session (for instance reminding the subjects to be careful during the annotation), and then immediately the subjects annotated the second half of the dialogue, and wrote up a summary. The second part of the second session was used for a separate experiment with a different dialogue and a slightly different annotation scheme.

4.2 The Annotation Scheme

MMAX 2 allows for multiple types of markables; markables at the phrase, utterance, and turn levels were defined before the experiment. All noun phrases except temporal ones were treated as phrase markables (Poesio, 2004). Subjects were instructed to go through the phrase markables in order (using MMAX 2's markable browser) and mark each of them with one of four attributes: "phrase" if it referred to an object which was mentioned earlier in the dialogue; "segment" if it referred to a plan, event, action, or fact discussed earlier in the dialogue; "place" if it was one of the five railway stations Avon, Bath, Corning, Dansville, and Elmira, explicitly mentioned by name; or "none" if it did not fit any of the above criteria, for instance if it referred to a novel object or was not a referential noun phrase. (We included the attribute "place" in order to avoid having our subjects mark pointers from explicit place names. These occur frequently in the dialogue–49 of the 151 markables–but are rather uninteresting as far as anaphora goes.) For markables designated as "phrase" or "segment" subjects were instructed to set a pointer to the antecedent, a markable at the phrase or turn level. Subjects were instructed to set more than one pointer in case of ambiguous reference. Markables which were not given an attribute or which were marked as "phrase" or "segment" but did not have an antecedent specified were considered to be data errors; data errors occurred in 3 out of the 151 markables in the dialogue, and these items were excluded from the analysis.

We chose to mark antecedents using MMAX 2's pointers, rather than its sets, because pointers allow us to annotate ambiguity: an ambiguous phrase can point to two antecedents without creating an association between them. In addition, MMAX 2 makes it possible to restrict pointers to a particular level. In our scheme, markables marked as "phrase" could only point to phrase-level antecedents while markables marked as "segment" could only point to turn-level antecedents, thus simplifying the annotation.

As in previous studies (Eckert and Strube, 2001; Byron, 2003), we only allowed a constrained form of reference to discourse segments: our subjects could only indicate turn-level markables as antecedents. This resulted in rather coarse-grained markings, especially when a single turn was long and included discussion of a number of topics. In a separate experiment we tested a more complicated annotation scheme which allowed a more fine-grained marking of reference to discourse segments.

4.3 Computing anaphoric chains

The raw annotation data were processed using custom-written Perl scripts to generate coreference chains and calculate reliability statistics.

The core of Passonneau's proposal (Passonneau, 2004) is her method for generating the set of dis-

[4] Available from `http://mmax.eml-research.de/`

tinct and mutually exclusive categories required by α out of the raw data of anaphoric annotation. Considering as categories the immediate antecedents would mean a disagreement every time two annotators mark different members of an anaphoric chain as antecedents, while agreeing that these different antecedents are part of the same chain. Passonneau proposes the better solution to view the emerging anaphoric chains themselves as the categories. And in a scheme where anaphoric reference is unambiguous, these chains are equivalence classes of markables. But we have a problem: since our annotation scheme allows for multiple pointers, these chains take on various shapes and forms.

Our solution is to associate each markable *m* with the set of markables obtained by following the chain of pointers from *m*, and then following the pointers backwards from the resulting set. The rationale for this method is as follows. Two pointers *to* a single markable never signify ambiguity: if *B* points to *A* and *C* points to *A* then *B* and *C* are cospecificational; we thus have to follow the links up and then back down. However, two pointers *from* a single markable may signify ambiguity, so we should not follow an up-link from a markable that we arrived at via a down-link. The net result is that an unambiguous markable is associated with the set of all markables that are cospecificational with it on one of their readings; an ambiguous markable is associated with the set of all markables that are cospecificational with at least one of its readings. (See figure 1.)

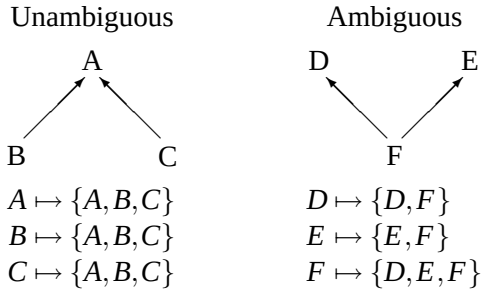


Figure 1: Anaphoric chains

This method of chain construction also allows to resolve apparent discrepancies between reference to phrase-level and turn-level markables. Take for example the snippet below: many annotators marked a pointer from the demonstrative *that* in utterance unit 4.2 to turn 3; as for *that* in utterance unit 4.3, some marked a pointer to the previous *that*, while others marked a pointer directly to turn 3.

(2)
```
3.1 M: and while it's there it
       should pick up the tanker
4.1 S: okay
4.2    and that can get
4.3    we can get that done by
       three
```

In this case, not only do the annotators mark different direct antecedents for the second *that*; they even use different attributes–"phrase" when pointing to a phrase antecedent and "segment" when pointing to a turn. Our method of chain construction associates both of these markings with the same set of three markables – the two *that* phrases and turn 3 – capturing the fact that the two markings are in agreement.[5]

4.4 Taking ambiguity into account

The cleanest way to deal with ambiguity would be to consider each item for which more than one antecedent is marked as denoting a set of interpretations, i.e., a set of anaphoric chains (Poesio, 1996), and to develop methods for comparing such sets of sets of markables. However, while our instructions to the annotators were to use multiple pointers for ambiguity, they only followed these instructions for phrase references; when indicating the referents of discourse deixis, they often used multiple pointers to indicate that more than one turn had contributed to the development of a plan. So, for this experiment, we simply used as the interpretation of markables marked as ambiguous the union of the constituent interpretations. E.g., a markable E marked as pointing both to antecedent A, belonging to anaphoric chain {A,B}, and to antecedent C, belonging to anaphoric chain {C,D}, would be treated by our scripts as being interpreted as referring to anaphoric chain {A,B,C,D}.

5 RESULTS

5.1 Agreement on category labels

The following table reports for each of the four categories the number of cases (in the first half) in which

[5]It would be preferable, of course, to get the annotators to mark such configurations in a uniform way; this however would require much more extensive training of the subjects, as well as support which is currently unavailable from the annotation tool for tracking chains of pointers.

a good number (18, 17, 16) annotators agreed on a particular label–phrase, segment, place, or none–or no annotators assigned a particular label to a markable. (The figures for the second half are similar.)

Number of judgments	18	17	16	0
phrase	10	3	1	30
segment			1	52
place	16	1	1	54
none	10	5	1	29

Table 2: Cases of good agreement on categories

In other words, in 49 cases out of 72 at least 16 annotators agreed on a label.

5.2 Explicitly annotated ambiguity, and its impact on agreement

Next, we attempted to get an idea of the amount of *explicit* ambiguity–i.e., the cases in which coders marked multiple antecedents–and the impact on reliability resulting by allowing them to do this. In the first half, 15 markables out of 72 (20.8%) were marked as explicitly ambiguous by at least one annotator, for a total of 55 explicit ambiguity markings (45 phrase references, 10 segment references); in the second, 8/76, 10.5% (21 judgments of ambiguity in total). The impact of these cases on agreement can be estimated by comparing the values of K and α on the antecedents only, before the construction of cospecification chains. Recall that the difference between the coefficients is that K does not allow for partial disagreement while α gives it some credit. Thus if one subject marks markable *A* as antecedent of an expression, while a second subject marks markables *A* and *B*, K will register a disagreement while α will register partial agreement. Table 3 compares the values of K and α, computed separately for each half of the dialogue, first with all the markables, then by excluding "place" markables (agreement on marking place names was almost perfect, contributing substantially to overall agreement). The value of α is somewhat higher than that of K, across all conditions.

5.3 Agreement on anaphora

Finally, we come to the agreement values obtained by using α to compare anaphoric chains computed as discussed above. Table 4 gives the value of α for the first half (the figures for the second half are similar). The calculation of α was manipulated under the following three conditions.

		With place	Without place
First Half	K	0.62773	0.50066
	α	0.65615	0.53875
Second Half	K	0.66201	0.44997
	α	0.67736	0.47490

The coefficient reported here as K is the one called K by Siegel and Castellan (1988).
The value of α is calculated using Passonneau's distance metric; for other distance metrics, see table 4.

Table 3: Comparing K and α

Place markables. We calculated the value of α on the entire set of markables (with the exception of three which had data errors), and also on a subset of markables – those that were not place names. Agreement on marking place names was almost perfect: 45 of the 48 place name markables were marked correctly as "place" by all 18 subjects, two were marked correctly by all but one subject, and one was marked correctly by all but two subjects. Place names thus contributed substantially to the agreement among the subjects. Dropping these markables from the analysis resulted in a substantial drop in the value of α across all conditions.

Distance measure. We used the three measures discussed earlier to calculate distance between sets: Passonneau, Jaccard, and Dice.[6]

Chain construction. Substantial variation in the agreement values can be obtained by making changes to the way we construct anaphoric chains. We tested the following methods.

NO CHAIN: only the immediate antecedents of an anaphoric expression were considered, instead of building an anaphoric chain.

PARTIAL CHAIN: a markable's chain included only phrase markables which occurred in the dia-

[6]For the nominal categories "place" and "none" we assign a distance of zero between the category and itself, and of one between a nominal category and any other category.

	With place markables			Without place markables		
	Pass	Jacc	Dice	Pass	Jacc	Dice
No chain	0.65615	0.64854	0.65558	0.53875	0.52866	0.53808
Partial	0.67164	0.65052	0.67667	0.55747	0.53017	0.56477
Inclusive [−top]	0.65380	0.64194	0.69115	0.53134	0.51693	0.58237
Exclusive [−top]	0.62987	0.60374	0.64450	0.49839	0.46479	0.51830
Inclusive [+top]	0.60193	0.58483	0.64294	0.49907	0.47894	0.55336
Exclusive [+top]	0.57440	0.53838	0.58662	0.46225	0.41766	0.47839

Table 4: Values of α for the first half of dialogue 3.2

logue before the markable in question (as well as all discourse markables).

FULL CHAIN: chains were constructed by looking upward and then back down, including all phrase markables which occurred in the dialogue either before or after the markable in question (as well as the markable itself, and all discourse markables).

We used two separate versions of the full chain condition: in the [+top] version we associate the top of a chain with the chain itself, whereas in the [−top] version we associate the top of a chain with its original category label, "place" or "none".

Passonneau (2004) observed that in the calculation of observed agreement, two full chains always intersect because they include the current item. Passonneau suggests to prevent this by excluding the current item from the chain for the purpose of calculating the observed agreement. We performed the calculation both ways – the inclusive condition includes the current item, while the exclusive condition excludes it.

The four ways of calculating α for full chains, plus the no chain and partial chain condition, yield the six chain conditions in Table 4. Other things being equal, Dice yields a higher agreement than Jaccard; considering both halves of the dialogue, the Passonneau measure always yielded a higher agreement that Jaccard, while being higher than Dice in 10 of the 24 conditions, and lower in the remaining 14 conditions.

The exclusive chain conditions always give lower agreement values than the corresponding inclusive chain conditions, because excluding the current item reduces observed agreement without affecting expected agreement (there is no "current item" in the calculation of expected agreement).

The [−top] conditions tended to result in a higher agreement value than the corresponding [+top] conditions because the tops of the chains retained their "place" and "none" labels; not surprisingly, the effect was less pronounced when place markables were excluded from the analysis. Inclusive [−top] was the only full chain condition which gave α values comparable to the partial chain and no chain conditions. For each of the four selections of markables, the highest α value was given by the Inclusive [−top] chain with Dice measure.

5.4 Qualitative Analysis

The difference between annotation of (identity!) anaphoric relations and other semantic annotation tasks such as dialogue act or wordsense annotation is that apart from the occasional example of carelessness, such as marking *Elmira* as antecedent for *the boxcar at Elmira*,[7] all other cases of disagreement reflect a genuine ambiguity, as opposed to differences in the application of subjective categories.[8]

Lack of space prevents a full discussion of the data, but some of the main points can already be made with reference to the part of the dialogue in (2), repeated with additional context in (3).

[7] According to our (subjective) calculations, at least one annotator made one obvious mistake of this type for 20 items out of 72 in the first half of the dialogue–for a total of 35 careless or mistaken judgment out of 1296 total judgments, or 2.7%.

[8] Things are different for associative anaphora, see (Poesio and Vieira, 1998).

(3)
```
1.4 M: first thing _I'd_ like you to do
1.5    is send engine E2 off with a boxcar
       to Corning to pick up oranges
1.6    uh as soon as possible
2.1 S: okay [6 sec]
3.1 M: and while it's there it
       should pick up the tanker
```

The two *it* pronouns in utterance unit 3.1 are examples of the type of ambiguity already seen in (1). All of our subjects considered the first pronoun a 'phrase' reference. 9 coders marked the pronoun as ambiguous between engine E2 and the boxcar, 6 marked it as unambiguous and referring to engine E2, and 3 as unambiguous and referring to the boxcar. This example shows that when trying to develop methods to identify ambiguous cases it is important to consider not only the cases of *explicit* ambiguity, but also so-called *implicit* ambiguity–cases in which subjects do not provide evidence of being consciously aware of the ambiguity, but the presence of ambiguity is revealed by the existence of two or more annotators in disagreement (Poesio, 1996).

6 DISCUSSION

In summary, the main contributions of this work so far has been (i) to further develop the methodology for annotating anaphoric relations and measuring the reliability of this type of annotation, adopting ideas from Passonneau and taking ambiguity into account; and (ii) to run the most extensive study of reliability on anaphoric annotation todate, showing the impact of such choices. Our future work includes further developments of the methodology for measuring agreement with ambiguous annotations and for annotating discourse deictic references.

ACKNOWLEDGMENTS

This work was in part supported by EPSRC project GR/S76434/01, ARRAU. We wish to thank Tony Sanford, Patrick Sturt, Ruth Filik, Harald Clahsen, Sonja Eisenbeiss, and Claudia Felser.

References

P. Buitelaar. 1998. *CoreLex : Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

D. Byron. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. Technical Report 703, University of Rochester.

M. Eckert and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

D. Gross, J. Allen, and D. Traum. 1993. The TRAINS 91 dialogues. TRAINS Technical Note 92-1, Computer Science Dept. University of Rochester, June.

L. Hirschman. 1998. MUC-7 coreference task definition, version 3.0. In N. Chinchor, editor, *In Proc. of the 7th Message Understanding Conference*.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

K. Krippendorff. 1980. *Content Analysis: An introduction to its Methodology*. Sage Publications.

C. D. Manning and H. Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

C. Müller and M. Strube. 2003. Multi-level annotation in MMAX. In *Proc. of the 4th SIGDIAL*.

M. Palmer, H. Dang, and C. Fellbaum. 2005. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*. To appear.

R. J. Passonneau. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript., December.

R. J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proc. of LREC*, Lisbon.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June.

M. Poesio, F. Bruneseaux, and L. Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker, editor, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

M. Poesio. 1996. Semantic ambiguity and perceived ambiguity. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*, chapter 8, pages 159–201. CSLI, Stanford, CA.

M. Poesio. 2004. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston, May.

E. F. Prince. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

A. Rosenberg and E. Binkowski. 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proc. of NAACL*.

C. L. Sidner. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

S. Siegel and N. J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill.

K. van Deemter and R. Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637. Squib.

B. L. Webber. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.

B. L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Annotating Discourse Connectives in the Chinese Treebank *

Nianwen Xue
Department of Computer and Information Science
University of Pennsylvania
xueniwen@linc.cis.upenn.edu

Abstract

In this paper we examine the issues that arise from the annotation of the discourse connectives for the Chinese Discourse Treebank Project. This project is based on the same principles as the PDTB, a project that annotates the English discourse connectives in the Penn Treebank. The paper begins by outlining range of discourse connectives under consideration in this project and examines the distribution of the explicit discourse connectives. We then examine the types of syntactic units that can be arguments to the discourse connectives. We show that one of the most challenging issues in this type of discourse annotation is determining the textual spans of the arguments and this is partly due to the hierarchical nature of discourse relations. Finally, we discuss sense discrimination of the discourse connectives, which involves separating discourse connective from non-discourse connective senses and teasing apart the different discourse connective senses, and discourse connective variation, the use of different connectives to represent the same discourse relation.

* I thank Aravind Johi and Martha Palmer for their comments. All errors are my own, of course.

1 Introduction

The goal of the Chinese Discourse Treebank (CDTB) Project is to add a layer of discourse annotation to the Penn Chinese Treebank (Xue et al., To appear), the bulk of which has also been annotated with predicate-argument structures. This project is focused on discourse connectives, which include *explicit connectives* such as subordinate and coordinate conjunctions, discourse adverbials, as well as *implicit discourse connectives* that are inferable from neighboring sentences. Like the Penn English Discourse Treebank (Miltsakaki et al., 2004a; Miltsakaki et al., 2004b), the CDTB project adopts the general idea presented in (Webber and Joshi, 1998; Webber et al., 1999; Webber et al., 2003) where discourse connectives are considered to be predicates that take abstract objects such as propositions, events and situations as their arguments. This approach departs from the previous approaches to discourse analysis such as the Rhetorical Structure Theory (Mann and Thompson, 1988; Carlson et al., 2003) in that it does not start from a predefined inventory of abstract discourse relations. Instead, all discourse relations are lexically grounded and anchored by a discourse connective. The discourse relations so defined can be *structural* or *anaphoric*. Structural discourse relations, generally anchored by subordinate and coordinate conjunctions, hold locally between two adjacent units of discourse (such as clauses). In contrast, anaphoric discourse relations are generally anchored by discourse adverbials and only one argument can be identified structurally in the local context while the other can only be de-

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 84–91,
Ann Arbor, June 2005.

rived anaphorically in the previous discourse. An advantage of this approach to discourse analysis is that discourse relations can be built up incrementally in a bottom-up manner and this advantage is magnified in large-scale annotation projects where inter-annotator agreement is crucial and has been verified in the construction of the Penn English Discourse Treebank (Miltsakaki et al., 2004a). This approach closely parallels the annotation of the the verbs in the English and Chinese Propbanks (Palmer et al., 2005; Xue and Palmer, 2003), where verbs are the anchors of predicate-argument structures. The difference is that the extents of the arguments to discourse connectives are far less certain, while the arity of the predcates is fixed for the discourse connectives.

This paper outlines the issues that arise from the annotation of Chinese discourse connectives, with an initial focus on explicit discourse connectives. Section 2 gives an overview of the different kinds of discourse connectives that we plan to annotate for the CDTB Project. Section 3 surveys the distribution of the discourse connectives and Section 4 describes the kinds of discourse units that can be arguments to the discourse connectives. Section 5 specifies the scope of the arguments of discourse relations and describes what should be included in or excluded from the text span of the arguments. Sections 6 and 7 describes the need for a mechanism to address sense disambiguation and discourse connective variation, drawing evidence from examples of explicit discourse connectives. Finally, Section 8 concludes this paper.

2 Overview of Chinese Discourse Connectives

With our theoretical disposition, a discourse connective is viewed as a predicate taking two abstract objects such as propositions, events, or situations as its arguments. A discourse connective can be either explicit or implicit. An explicit discourse connective is realized in the form of one lexical item or several lexical items while an implicit discourse connective must be inferred between adjacent discourse units. Typical explicit discourse connectives are subordinate and coordinate conjunctions as well as discourse adverbials. While the arguments for subordinate and coordinate conjunctions are generally local, the first argument for a discourse adverbial may need to be identified long-distance in the previous discourse.

2.1 Subordinate conjunctions

There are two types of subordinate conjunctions in Chinese, single and paired. With single subordinate conjunctions, the subordinate conjunction introduces the subordinate clause, as in (1). By convention, the subordinate clause is labeled *ARG1* and the main clause is labeled *ARG2*. The subordinate conjunction is NOT included as part of the argument. The subordinate clause generally precedes the main clause in Chinese, but occasionally it can also follow the main clause. The assignment of the argument labels to the discourse units is independent of their syntactic distributions. The subordinate clause is always labeled *ARG1* whether it precedes or follows the main clause.

Simple subordinate conjunctions: Simple subordinate conjunctions are very much like English where the subordinate clause is introduced by a subordinate conjunction:

(1) 报告 认为 ，[*conn* 如果] [*arg1* 经济 和
report believe , **if** economic and
金融 政策 得力]，[*arg2* 亚洲地区 经济
financial policy effective , Asia region economy
可望 在１９９９年开始 回升]。
expect in 1999 begin recover .

”The report believes that if the economic and financial policies are effective, Asian economy is expected to recover in 1999.”

Paired subordinate conjunctions: Chinese also abounds in paired subordinate conjunctions, where the subordinate conjunction introduces the subordinate clause and another discourse connective introduces the main clause, as in (2). In this case, the discourse connectives are considered to be paired and jointly anchor ONE discourse relation.

(2) [*conn* 如果] [*arg1* 改革 措施 不 得力 ，
if reform measure not effective ,
信心 危机 依然存在]，[*conn* 那么] [*arg2*
confidence crisis still exist , **then**
投资者 就 有 可能 把 注意力 转向其他
investor will have possibility BA attention turn other
新兴 市场]。
emerging market .

”If the reform measures are not effective, confidence crisis still exists, then investors is likely to turn their attention to other emerging markets.”

Modified discourse connectives: Like English, some subordinate conjunctions can be modified by an adverb, as illustrated in (3). Note that the subordinate conjunction is in clause-medial position. When this happens, the first argument, ARG1 in this case, becomes discontinuous. Both portions of the argument, the one that comes before the subordinate conjunction and the one after, are considered to be part of the same argument.

(3) [*arg1* 去年 初 浦东 新区 诞生 的
last year beginning Pudong new district open DE
中国 第一家 医疗 机构 药品采购 服务
China first CL medical institution drug purchase service
中心]，[*conn* 正 因为] [*arg1* 一 开始 就
center , **just because** once begin
比较 规范]，[arg2 运转 至今 ，成交
relatively standardized , operate till now , trade
药品 一亿多 元 ，没有发现一 例
medicine over 100 million yuan , not find one case
回扣]。
killback .

"It is because its operations are standardized that the first purchase service center for medical institutions in China opened in the new district of Pudong in the beginning of last year has not found a single case of kickback after it has traded 100 million yuan worth of medicine in its operation till now."

Conjoined discourse connectives: The subordinate conjunctions can be conjoined in Chinese so that there are two subordinate clauses each having one instance of the same subordinate conjunction. In this case, there is still one discourse relation, but ARG1 is the conjunction of the two subordinate clauses. This is in contrast with English, where only one subordinate conjunction is possible and ARG1 is linked with a coordinate conjunction, as illustrated in the English translation.

(4) [conn 虽然] [arg1 黄春明 已经
although Huang Chunming already
十几 年 没有出版 小说集 了]，[conn
over 10 year not publish novel series AS ,
虽然] [*arg2* 从 〈城仔 落 车 〉到〈
although from " city boys miss bus " to "
售票口 〉，中间 隔 了 三十七 年]，
ticket box ", middle span AS thirty seven year ,
[conn 但] [arg2 黄春明 的 文学 内在 ，
but Huang Chunming DE literary theme ,
有些 东西 竟然 从来都 没有改变]。
some thing surprisingly ever have not change .

"Although Huang Chunming has not published a novel series for over ten years, and it spans over thirty seven years from 'City Boys Missed Bus' to 'Ticket Box', surprisingly some things in Huang Chunming's literary themes have never changed."

2.2 Coordinate conjunctions

The second type of explicit discourse connectives we annotate are coordinate discourse conjunctions. The arguments of coordinate conjunctions are annotated in the order in which they appear. The argument that appears first is labeled ARG1 and the argument that appears next is marked ARG2. The coordinate conjunctions themselves, like subordinate conjunctions, are excluded from the arguments.

(5) 近年 来，美国 每 年 糖尿病
recent years in , the U.S. every year diabetes
医疗费 约 一百亿 美元 ，印度 去年
medical expense about 10 billion dollar , India last year
糖尿病 医疗费 为
diabetes medical expenses be
六点一亿 美元 ，[*arg1* 中国 尚
six hundred and 10 million dollar , China yet
无 具体 统计]，[*conn* 但] [*arg2* 中国
not have concrete statistics , **but** China
糖尿病 人数 正 以 每 年 七十五万
diabetes population currently with every year 750,000
新 患者 的 速度 递增]。
new patient DE speed increase .

"In recent years, the medical expenses for diabetes patients in the U.S. is about 10 billion dollars. Last year the medical expenses for diabetes patients in India is six hundred and ten million dollars. China does not have concrete statistics yet, but its diabetes population is increasing at a pace of 750,000 new patients per year.

Paired coordinate conjunctions: Like subordinate conjunctions, coordinate conjunctions can also be paired, as in (6):

(6) 现代 父母 难 为的 地方 在于 [*conn* 既
modern parent difficult be DE place lie in **CONN**
] [*arg1* 无法 排除 血液 中流传的 观念]，
no way eliminate blood in flow DE tradition ,
[*conn* 又] [*arg2* 要 面对新 的 价值]。
CONN need face new DE value .

"The difficulty of being modern parents lies in the fact they can not get rid of the traditional values flowing in their blood, and they also need to face new values."

2.3 Adverbial connectives

The third type of explicit discourse connectives we annotate are discourse adverbials. A discourse adverbial differs from other adverbs in that they require an antecedent that is a proposition or a set of related propositions. Generally, the second argument is adjacent to the discourse adverbial while the first argument may be long-distance. By convention, the second argument that is adjacent to the discourse connective is labeled ARG2 and the other argument is

marked as ARG1. Note that in (7b) that first argument is not adjacent to the discourse adverbial.

(7) a. 美国 商会 广东
The U.S. Chamber of Commerce Guangdong
分会 会长 康永华 律师 说 ,
Chapter Chairman Kang Yonghua lawyer say ,
[*arg1* 克林顿 政府 已经 表示 要
Clinton Administration already indicate will
延长 中国 的 贸易 最惠国待遇] , [*conn*
renew China DE trade MFN status ,
因此] , [*arg2* 这 次 游说 的 重点 是
therefore , this time lobby DE focus be
那些 较 保守 的 议员] 。
those relatively conservative DE congressman .

"Lawyer Kang Yonghua, chairman of the Guangdong Chapter of the U.S. Chamber of Commerce, says that since the Clinton Administration has already indicated that it will renew China's MFN status, the focus of the lobby this time is on those relatively conservative congressmen."

b. [*arg1* 中国 批准 的 外企 中 ,
China approve DE foreign enterprise in ,
工业 项目 占 七成
industry project account for seventy percent,
, 其中 加工 工业偏 多
among them processing industry excessive
], 这 与 中国 劳动力 素质 、成本
, this with China labor force training , cost
较 低 的 国情 相吻合 ,
relatively low DE state of affairs consistent ,
[*conn* 从而] [*arg2* 吸纳 了 大量
therefore absorb ASP big volume
劳动力] 。
labor force .

"In the foreign enterprises that China approved of, industry projects accounts for seventy percent of them. Among them processing projects are excessively high. This is consistent with the current state of affairs in China where the training and cost of the labor force is low. Therefore they absorbed a large portion of the labor force."

2.4 Implicit discourse connectives

In addition to the explicit discourse connectives, there are also implicit discourse connectives that must be inferred from adjacent propositions. The arguments for implicit discourse connectives are marked in the order in which they occur, with the argument that occurs first marked as ARG1 and the other argument marked as ARG2. By convention a punctuation mark is reserved as the place-holder for the discourse connective. Where possible, the annotator is asked to provide an explicit discourse connective to characterize the type of discourse relation. In (8), for example, a coordinate conjunction 而"while" can be used in the place of the implicit discourse connective.

(8) [*arg1* 其中 出口 为 一百七十八点三亿美元
among them export be 17.83 billion dollar
, 比 去年 同 期 下降
, compared with last year same period decrease
百分之一点三] [**conn=**而 ;] [*arg2* 进口
1.3 percent ; import
一百八十二点七亿美元 , 增长
18.27 billion dollar , increase
百分之三十四点一] 。
34.1 percent .

"Among them, export is 17.83 billion, an 1.3 percent increase over the same period last year. Meanwhile, import is 18.27 billion, which is a 34.1 percent increase."

3 Where are the discourse connectives?

In Chinese, discourse connectives are generally clause-initial or clause-medial, although localizers are clause-final and can be used as discourse connective by themselves or together with a preposition. Subordinate conjunctions, coordinate conjunctions and discourse adverbial can all occur in clause-initial as well as clause-medial positions. The distribution of the discourse connectives is not uniform, and varies from discourse connective to discourse connective. Some discourse connectives alternate between clause-initial and clause-medial positions. The examples in (9) show that 尽管"even though", which forms a paired connective with 但是"but", occurs in both clause-initial (9a) and clause-medial (9b) positions.

(9) a. [*conn* 尽管] [*arg1* 亚洲 一些 国家 的
even though Asia some country DE
金融 动荡 会 使 这些 国家 的
financial turmoil will make these country DE
经济 增长 受到 严重 影响] ,
economy growth experience serious impact ,
[*conn* 但] [*arg2* 就整 个 世界 经济 而言
but to whole CL world economy
, 其他 国家 的 强劲 增长 势头 会
, other country DE strong growth momentum will
弥补 这 一 损失] 。
compensate this one loss .

"Even though the financial turmoil in some Asian countries will affect the economic growth of these countries, as far as the economy of the whole world is concerned, the strong economic growth of other countries will make up for this loss."

b. [*arg1* 展望 虎年 , 中国 的
look ahead Year of Tiger , China DE
经济 列车] [*conn* 尽管] [*arg1* 会
economy train **even though** will

有 颠簸 起伏]，[*conn* 但] [*arg2* 只要
have ups and downs , **but** as long as
调控 措施 适时 、得当 ，相信 会 沿着
adjust measure timely , proper , believe will along
预设 的 轨道 稳健 前行]。
expect DE track steady advance .

"Looking ahead at the Year of Tiger, even though China's economic train will have its ups and downs, as long as the adjusting measures are timely and proper, we believe that it will advance steadily along the expected track."

Localizers are a class of words that occur after clauses or noun phrases to denote temporal or spatial discourse relations. They can introduce a subordinate clause by themselves or together with a preposition. While the preposition is optional, the localizer is not. When both the preposition and the localizer occur, they form a paired discourse connective anchoring a discourse relation. Example (10) shows the preposition 当and the localizer 时form a paired discourse connective equivalent to the English subordinate conjunction "when".

(10) 日前 ，[*conn* 当] [*arg1* 记者 在这里
a few days ago , **when** reporter at here
专访 欧盟欧洲 委员会 驻华
interview exclusively EU Europe Commission to China
代表团 团长魏根深 大使 ，请 他
delegation head Wei Genshen ambassador , ask he
评价 这 一 年 来 双方 的 合作
comment this one year since two sides DE cooperation
成果] [**conn** 时]，[*arg2* 他 毫 不
accomplishment **when** , he little no
迟疑 地 说 ： "欧盟同 中国 的 政治
hesitate DE say : ' EU with China DE political
关系 、贸易 关系 以及在投资 等 方面
relation , trade relation and at investment etc. aspect
的 合作 在一九九七年都 取得 了
DE cooperation in 1997 all achieve ASP
显著 的 发展 。"]
significant DE progress . "

"A few days ago, when this reporter exclusively interviewed Wei Genshen, head of the EU Europe Commission delegation to China, and asked him to comment on the accomplishment of the cooperation between the two sides in the past year, without any hesitation he said: 'There was significant progress in the political relations, trade relations, and the cooperation in trade, etc. between EU and China.' "

4 What counts as an argument?

This section examines the syntactic composition of arguments to discourse connectives in Chinese. Arguments of discourse relations are propositional situations such as events, states, or properties. As such an argument of a discourse relation can be realized as a clause or multiple clauses, a sentence or multiple sentences. Typically, a subordinate conjunction introduces clauses that are arguments in a discourse relation. Discourse adverbials and coordinate conjunctions, however, can take one or more sentences to be their arguments. The examples in (11) shows that arguments to discourse connectives can be a single clause (11a), multiple clauses (11b), a single sentence (11c) and multiple sentences (11d) respectively.

(11) a. [*conn* 尽管] [*arg1* 今年 一 至
even though this year January to
十一月 中国 批准 利用 外资
November China approve utilize foreign investment
项目 数 和 合同 外资
project number and contract foreign investment
金额 都 比 去年 同 期
amount both compared with last year same period
有所下降]，[*conn* 但] [*arg2* 实际 利用
have decrease , **but** actually use
外资 金额 仍 比
foreign investment amount still compared with
去年 同 期 增长 了
last year same period increase ASP
百分之二十七点零一] 。
27.01 percent .

"Even though the number of projects that use foreign investment that China approved of and contractual foreign investment both decreased compared with the same period last year, the foreign investment that has actually been used increased 27.01 percent."

b. [*conn* 由于] [*arg1* 茅台酒 制作 工艺
because Maotai Liquor brew process
复杂 ，生产 周期 长]，[*conn*
complicated , production cycle long ,
因而] [*arg2* 其 产量 十分有限
therefore its production volume very limited
] 。
.

"Because the brewing process of Maotai liquor is complicated and its production cycle is long, its production volume is very limited."

c. [*arg1* 中国 乒乓球 运动员没有参加
Chinese table tennis athlete not participate
第二十九 和 三十 届
twenty-ninth and thirtieth CL
世乒赛] 。[*conn* 因此]
word table tennis tournament . **therefore**
，[*arg2* 复制 的 金牌 中包括 将 要
, replicate DE gold medal in include will will
举行的 第四十五 届
hold DE forty-fifth CL
世乒赛 金牌] 。
world table tennis tournament gold medal .

”Chinese athletes did not attend the twenty-ninth and the thirtieth world table tennis tournaments. Therefore, The replicated gold medals also include the gold medals in the yet-to-be-held forty-fifth world tournament.”

d. [*arg1* 回归 后 对 澳门 的 未来 发展 是
return after for Macao DE future prospect be
利 还是弊 ？有 五成三 的 人
plus or minus ? have 53 percent DE people
回答 不 知道]。[*conn* 但] [*arg2* 对于能 不
answer not know . **but** to can not
能 接受 和 港 澳 一样，以 「
can accept like Hong Kong Macao same , with ’
一 国 两 制 」解决 台湾 问题，则
one country two system ’ resolve Taiwan issue ,
有 二成七 的 民众 表示 「不 知道 」
have 27 percent DE people indicate ’ not know ’
，五成九 的 民众 表示 「不 能 接受 」
, 59 percent DE people indicate ’ not can accept ’
] 。
.

”Is the return of sovereignty (to China) a plus or minus for Macao’s future? 53 percent of people say they don’t know. But to the question of whether they accept the resolution of the Taiwan issue with ’one country, two systems’ like Hong Kong and Macao, 59 percent of the people say ’they cannot accept’ . ”

5 Argument Scope

Determining the scope of an argument to a discourse connective has proved to be the most challenging part of the discourse annotation. A lot of the effort goes into deciding when certain text units should be included in or excluded from the argument of a discourse connective. Under our annotation scheme, the prepositional phrases, which generally precede the subject in a Chinese clause, are included in the argument of a discourse connective, as illustrated in (12a). The material in the main clause that embeds a discourse relation, however, are excluded, as in (12b).

(12) a. 另外 ，[*arg1* 在休闲 文化 生活缺乏
in addition , in recreation culture life lack
的 东莞]，[*conn* 除非] [*arg1* 很 有
DE Dongguan , **unless** very have
教育 热诚]，[*conn* 否则] [*arg2*
education enthusiasm , **otherwise**
很 难 留住教师]。
very difficult keep teacher .

” In addition, in Dongguan where recreational activities are lacking, unless they are very enthusiastic about education, it is very hard to keep teachers.”

b. 任志刚 还 表示 ，[*conn* 由于] [*arg1*
Ren Zhigang also indicate , **because**
香港 和 美国 息 差 达
Hong Kong and the U.S. interest discrepancy reach
一百二十五点]，[*arg2* 如果市场 对
125 point , if market in
香港 经济 前景 充满 信心 ，
Hong Kong economic prospect full of confidence ,
仍 有 减 息 空间]。
still have reduce interest space .

”Ren Zhigang also indicated that because the interest discrepancy between Hong Kong and the U.S. reaches 125 point, if the market is fully confident in the economic prospect of Hong Kong, there is still room for reducing interest rates.”

A lot of the challenge in determining the scope of an argument stems from the fact that discourse structures are recursive. As such identifying the scope of an argument is effectively determining how the discourse relations are hierarchically organized. This is illustrated in (13), where the discourse relation anchored by the coordinate conjunction 但”but” is embedded within the discourse relation anchored by the subordinate conjunction 如果”if”. The ambiguity is whether the conditional clause introduced by ”如果” has scope over one or two of the clauses coordinated by 但”but”.

(13) 报告 认为 ，[**conn** 如果] [**arg1** 经济 和
report believe , **if** economy and
金融 政策 得力]，[**arg2** [*arg1* 亚洲地区
finance policy effective , Asia region
经济 可望 在１９９９年开始 回升]，[*conn*
economy expect in 1999 begin recover ,
但] [*arg2* 不 会 象 墨西哥和 阿根廷 在
but not will like Mexico and Argentina in
１９９４－１９９５年金融 危机后 那样
1994 to 1995 finance crisis after like that
出现 高速 Ｖ形 大 回升]]。
occur high-speed V-shaped big recovery .

”The report believes that if the economic and financial policies are effective, the economy of Asia is expected to recover, but there will not be a V-shaped high-speed recovery like the one after the financial crisis of Mexico and Argentina in 1994 and 1995.”

Given our bottom-up approach in which discourse connectives anchor binary discourse relations, we do not explicitly annotate hierarchical structures between the arguments. However, such discourse relations can be deduced when some discourse relations are recursively embedded within another as arguments to another discourse connective.

6 Sense Disambiguation

Although discourse connectives are often considered to be a closed set, some lexical items in Chinese can be used as both a discourse connective and a non-discourse connective. In this case it is important to tease them part. There are also discourse connectives that have different senses, and it is potentially beneficial for certain NLP applications to disambiguate these senses. Machine Translation, for example, would need to translate the different senses into different discourse connectives in the target language. The examples in (14) shows the different senses of 而, which can be translated into "while" (14a), "but" (14c), "and" (14d) and "instead" (14e). Note that in (14e) it is important for the first argument to be negated by 不"not". In (14b), however, it is not a discourse connective. It does not seem to contribute any meaning to the sentence and is probably just there to satisfy some prosodic constraint.

(14) a. 1997年发达 国家 经济 形势
1997 developed country economic situation
的 特点 是 [*arg1* 美国 增长 强劲]
DE characteristic be U.S. grow strongly
[*conn* 而] [*arg2* 日本 经济 疲软] , 美国
while Japan economy weak , U.S.
经济 增长率 估计 为 百分之三点七 ,
economic growth estimate be 3.7 percent ,
日本 仅 为 百分之零点八 。
Japan only be 0.8 percent .

"The economic situation in developed countries in 1997 is that the U.S. (economy) grows strongly while the Japanese economy is weak. The U.S. economic growth rate was estimated to be 3.7 percent while the Japanese economy grows at 0.8 percent."

b. 水东 开发区 位 于
Shuidong Development Zone located at
粤西 地区 的 茂名市
western Guangdong region DE Maoming city
境内 , 面积 八十多 平方公里 ,
territory , coverage over eighty square kilometer ,
是 适应 乙烯 工程 的 需要 [? 而] 建立
be suit ethylene project DE need **?** establish
的 一 个 后继 加工 基地 。
DE one CL downstream process base .

"Shuidong Development Zone, located in Maoming City of western Guangdong occupies an area of over eighty square kilometers. It is a downstream processing base established to meet the need of the ethylene project."

c. 能 生产 [*arg1* 中国 不 能 生产] [*conn*
can produce China not can produce
而] [*arg2* 又 很 需要] 的 药品 的
but again badly need DE drug DE
企业
enterprise

"Enterprises that can produce drugs that China badly needs but cannot produce"

d. 吉林省 珲春市 市长 金硕仁 说
Jilin Province Huichun City mayor Jin Shuoren say
: "国际 社会 的 支持 和
: " international community DE support and
参与 , 对于 珲春 的 开发
participation , to Huichun DE development
开放 起 了 [*arg1* 积极]
opening to the outside play DE positive
[*conn* 而] [*arg2* 关键] 的作用 。"
and key DE role . "

"Jing Shuoren, mayor of Huichun City of Jilin Province said: "The support and participation of the international community played a positive and key role in Huichun's development and opening up to the outside."

e. [*arg1* 这 当然 不 是 历史 的 巧合]
this certainly not be history DE coincidence
, [*conn* 而] [*arg2* 是 历史 的
, **instead** be history DE
积累 和 转接] 。
accumulation and transition .

"This certainly is not historical coincidence. Instead it is historical accumulation and transition."

7 Discourse Connective Variation

The flip side of sense disambiguation is that one discourse relation is often realized with different discourse connectives due to the long evolution of the Chinese language and morphological processes like *suoxie*, which is one form of abbreviation. The examples in (15) shows the different variations of the discourse relation of concession. The different forms of the discourse connective are so similar that they can hardly be considered to be different discourse connectives. In principle, any combination of part 1 and part 2 from Table 7 can form a paired discourse connective, subject to some non-discourse related constraints. In (15a), for example, the abbreviated 虽 can only occur in clause-medial positions. (15b) shows the second part of the paired discourse connective can be dropped without changing the semantics of the discourse relation. (15c) shows that the second part of the paired discourse connective can be combined with another discourse connective.

(15) a. [*arg1* 王翔] [*conn* 虽] [*arg1*
Wang Xiang **although**
年过半百] , [*conn* 但] [*arg2* 其
over fifty years old , **but** his

gloss	discourse connectives
although	[1] 虽然, 虽说, 虽 [2] 但是, 但, 还是, 可是, 却, 然而, 不过
because	[1] 因为, 因, 由于 [2] 所以
if	[1] 如果, 若, 假如 [2] 就
therefore	因此, 于是

Table 1: Discourse connective variation

充沛 的 精力 和 敏捷 的 思维 ，给
abundant DE energy and quick DE thinking , give
人 以 一 个 挑战者 的 印象]。
people with one CL challenger DE impression .

"Although Wang Xiang is over fifty years old, but his abundant energy and quick thinking gives people the impression of a challenger."

b. [*arg1* 外在 的 环境] [*conn* 虽然]
external DE environment **although**
[*arg1* 改变 了]，[*arg2* 内心 那 份 渴望
change ASP , heart that CL long for
记忆 与 归属 的 需求 很
memory and sense of belonging DE need very
难 改变]。
difficult change .

"Although the external environment has changed, the need of longing for memory and sense of belonging is very difficult to change."

c. [*arg1* 大陆 政策] [*conn* 虽然] [*arg1*
mainland policy **although**
动辄得咎]，[*conn* 但 却] [*arg2* 是
vulnerable to criticism , **but but** be
所有政策 的 基础]，任何候选人 都 无法
all policy DE basis , any candidate all cannot
忽视 。
ignore .

"Although the mainland policy is vulnerable to criticism, it is the basis of all policies and no candidate afford to ignore it."

8 Conclusion

We examined the range of discourse connective we plan to annotate for the Chinese Discourse Treebank project. We have shown that while arguments to subordinate and coordinate conjunctions can be identified locally, arguments to discourse adverbials may be long-distance. We also examined the distribution of the discourse connectives in Chinese and the syntactic composition and the scope of the arguments in discourse relations. We have shown the most challenging issue in discourse annotation is determining the text span of a discourse argument and this is partly due to the hierarchical nature of discourse structures. We have discussed the need to address sense disambiguation and discourse connective variation in our annotation of Chinese discourse connectives.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

William Mann and Sandra Thompson. 1988. Rhetorical Sturcture Theory. *Text*, 8(3):243–281.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004a. The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004b. The Penn Discourse Treebank. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston, Massachusetts.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

B. Webber and A. Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In *In ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalized TAG. In *Meeting of the Association of Computational Linguistics*, College Park, MD.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. To appear. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*.

Author Index