Discourse Structure and Computation: Past, Present and Future

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh UK EH8 9AB
`bonnie.webber@ed.ac.uk`

Aravind Joshi
Dept of Computer & Information Science
University of Pennsylvania
Philadelphia PA 19104-6228
`joshi@seas.upenn.edu`

Abstract

The discourse properties of text have long been recognized as critical to language technology, and over the past 40 years, our understanding of and ability to exploit the discourse properties of text has grown in many ways. This essay briefly recounts these developments, the technology they employ, the applications they support, and the new challenges that each subsequent development has raised. We conclude with the challenges faced by our current understanding of discourse, and the applications that meeting these challenges will promote.

1 Why bother with discourse?

Research in Natural Language Processing (NLP) has long benefitted from the fact that text can often be treated as simply a bag of words or a bag of sentences. But not always: *Position* often matters — e.g., It is well-known that the first one or two sentences in a news report usually comprise its best extractive summary. *Order* often matters – e.g., very different events are conveyed depending on how clauses and sentences are ordered.

(1) a. I said the magic words, and a genie appeared.

 b. A genie appeared, and I said the magic words.

Adjacency often matters — e.g., attributed material may span a sequence of adjacent sentences, and contrasts are visible through sentence juxtaposition. *Context* always matters — e.g., All languages achieve economy through minimal expressions that can only convey intended meaning when understood in context.

Position, order, adjacency and context are intrinsic features of *discourse*, and research on discourse processing attempts to solve the challenges posed by context-bound expressions and the discourse structures that give rise, when linearized, to position, order and adjacency.

But challenges are not why Language Technology (LT) researchers should care about discourse: Rather, discourse can enable LT to overcome known obstacles to better performance. Consider automated summarization and machine translation: Humans regularly judge output quality in terms that include *referential clarity* and *coherence*. Systems can only improve here by paying attention to discourse — i.e., to linguistic features above the level of n-grams and single sentences. (In fact, we predict that as soon as cheap — i.e., non-manual – methods are found for reliably assessing these features — for example, using proxies like those suggested in (Pitler et al., 2010) — they will supplant, or at least complement today's common metrics, *Bleu* and *Rouge* that say little about what matters to human text understanding (Callison-Burch et al., 2006).)

Consider also work on *automated text simplification*: One way that human editors simplify text is by re-expressing a long complex sentence as a discourse sequence of simple sentences. Researchers should be able to automate this through understanding the various ways that information is conveyed in discourse. Other examples of LT applications already benefitting from recognizing and applying discourse-level information include automated assessment of student essays (Burstein and Chodorow, 2010); summarization (Thione et al., 2004), infor-

mation extraction (Patwardhan and Riloff, 2007; Eales et al., 2008; Maslennikov and Chua, 2007), and more recently, statistical machine translation (Foster et al., 2010). These are described in more detail in (Webber et al., 2012).

Our aim here then, on this occasion of ACL's 50th Annual Meeting, is to briefly describe the evolution of computational approaches to discourse structure, reflect on where the field currently stands, and what new challenges it faces in trying to deliver on its promised benefit to Language Technology.

2 Background

2.1 Early Methods

The challenges mentioned above are not new. Question-Answering systems like LUNAR (Woods, 1968; Woods, 1978) couldn't answer successive questions without resolving context-bound expressions such as pronouns:

 (2) What is the concentration of silicon in breccias?
 ⟨*breccia1*, *parts per million*⟩
 ⟨*breccia2*, *parts per million*⟩
 ⟨ . . . ⟩
 What is *it* in volcanics? (Woods, 1978)

Early systems for human interaction with animated agents, including SHRDLU (Winograd, 1973) and HOMER (Vere and Bickmore, 1990), faced the same challenge. And early message understanding systems couldn't extract relevant information (like when a sighted submarine submerged – "went sinker") without recognizing relations implicit in the structure of a message, as in

 (3) VISUAL SIGHTING OF PERISCOPE FOLLOWED BY ATTACK WITH ASROC AND TORPEDO. WENT SINKER. LOOSEFOOT 722/723 CONTINUE SEARCH. (Palmer et al., 1993)

The same was true of early systems for processing narrative text (under the rubric *story understanding*). They took on the problem of recognizing events that had probably happened but hadn't been mentioned in the text, given the sequence of events that had been (Lehnert, 1977; Rumelhart, 1975; Schank and Abelson, 1977; Mandler, 1984).

Since these early systems never saw more than a handful of examples, they could successfully employ straight-forward, but *ad hoc* methods to handle

the discourse problems the examples posed. For example, LUNAR used a single 10-position ring buffer to store discourse entities associated with both the user's and the system's referring expressions, resolving pronouns by looking back through the buffer for an appropriate entity and over-writing previous buffer entries when the buffer was full.

The next wave of work in computational discourse processing sought greater generality through stronger theoretical grounding, appealing to then-current theories of discourse such as *Centering Theory* (Grosz et al., 1986; Grosz et al., 1995), used as a basis for anaphor resolution (Brennan et al., 1987; Walker et al., 1997; Tetreault, 2001) and text generation (Kibble and Power, 2000), *Rhetorical Structure Theory* (Mann and Thompson, 1988), used as a basis for text generation (Moore, 1995) and document summarization (Marcu, 2000b), and Grosz and Sidner's theory of discourse based on intentions (Grosz and Sidner, 1986a) and shared plans (Grosz and Sidner, 1990), used in developing animated agents (Johnson and Rickel, 2000). Issues related to fully characterizing centering are explored in great detail in (Kehler, 1997) and (Poesio et al., 2004).

The approaches considered during this period never saw more than a few handfuls of examples. But, as has been clear from developments in PoS-tagging, Named Entity Recognition and parsing, Language Technology demands approaches that can deal with whatever data are given them. So subsequent work in computational discourse processing has similarly pursued robustness through the use of data-driven approaches that are usually able to capture the most common forms of any phenomenon (ie, the 80% at the high end of the Zipfian distribution), while giving up on the long tail. This is described in Section 3.

2.2 Early Assumptions

While early work focussed on the correct assumption that much was implicit in text and had to be inferred from the explicit sequence of sentences that constituted a text, work during the next period focussed on the underlying structure of discourse and its consequences. More specificaly, it assumed that the sequence of sentences constituting the text were covered by a single tree structure, similar to the single tree structure of phrases and clauses covering the
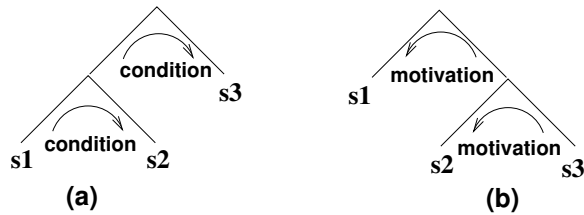
Figure 1: Proposed discourse structures for Ex. 4: (a) In terms of informational relations; (b) in terms of intentional relations

words in a sentence. At issue though was the nature of the structure.

One issue concerned the nature of the relation between parent and child nodes in a discourse tree, and/or the relation between siblings. While *Rhetorical Structure Theory* (Mann and Thompson, 1988) posited a single discourse relation holding between any two *discourse units* (i.e., units projecting to adjacent text spans), Moore and Pollack (1992) gave an example of a simple discourse (Ex. 4) in which different choices about the discourse relation holding between pairs of units, implied different and non-isomorphic structures.

(4) Come home by 5:00. $_{s1}$ Then we can go to the hardware store before it closes. $_{s2}$ That way we can finish the bookshelves tonight. $_{s3}$

Example 4 could analysed purely in terms of information-based discourse relations, in which $s1$ specified the CONDITION under which $s2$ held, which in turn specified the CONDITION under which $s3$ held. This would make $s1$ subordinate to $s2$, which in turn would be subordinate to $s3$, as in Figure 1a. Alternatively, Example 4 could be analysed purely in terms of intention-based (pragmatic) relations, in which $s2$ would be MOTIVATION for $s1$, while $s3$ would be MOTIVATION for $s2$. This would make $s3$ subordinate to $s2$, which in turn would be subordinate to $s1$, as in Figure 1b. In short, the choice of relation was note merely a matter of labels, but had structural implications as well.

Another issue during this period concerned the nature of discourse structure: Was it really a tree? Sibun (1992), looking at people's descriptions of the layout of their house or apartment, argued that they resembled different ways of linearizing a graph of the rooms and their connectivity through doors and halls. None of these linearizations were trees. Similarly, Knott *et al.* (2001), looking at transcriptions of museum tours, argued that each resembled a linear sequence of trees, with one or more topic-based connections between their root nodes — again, not a single covering tree structure. Wiebe (1993), looking at simple examples such as

(5) The car was finally coming toward him. $_{s1}$
He finished his diagnostic tests, $_{s2}$
feeling relief. $_{s3}$
But then the car started to turn right. $_{s4}$

pointed multiple lexical items explicitly relating a clause to multiple other clauses. Here, but would relate $s4$ to $s3$ via a CONTRAST relation, while then would relate $s4$ to $s2$ via a temporal SUCCESSION relation.

The most well-known of work from this period is that of Mann and Thompson (1988), Grosz and Sidner (1986b), Moore and Moser (1996), Polanyi and van den Berg (1996), and Asher and Lascarides (2003).[1]

The way out of these problems was also a way to achieve the robustness required of any Language Technology, and that lay in the growing consensus towards the view that discourse does not have a single monolithic hierarchical structure. Rather, different aspects of a discourse give rise to different structures, possibly with different formal properties (Stede, 2008; Stede, 2012; Webber et al., 2012). These different structures we describe in the next section, while the fact that this can't be the end of the story, we take up in Section 4.

3 The Situation Today

Recent years have seen progress to differing degrees on at least four different types of discourse structures: topic structure, functional structure, event structure, and a structure of coherence relations. First we say a bit about the structures, and then about the resources employed in recognizing and labelling them.

3.1 Types of discourse structures

Topic structure and *automated topic segmentation* aims to break a discourse into a linear sequence of

---

[1] For a historical account and assessent of work in automated anaphora resolution in this period and afterwards, we direct the reader to Strube (2007), Ng (2010) and Stede (2012).

topics such the geography of a country, followed by its history, its demographics, its economy, its legal structures, etc. Segmentation is usually done on a sentence-by-sentence basis, with segments not assumed to overlap. Methods for topic segmenation emply semantic, lexical and referential similarity or, more recently, language models (Bestgen, 2006; Chen et al., 2009; Choi et al., 2001; Eisenstein and Barzilay, 2008; Galley et al., 2003; Hearst, 1997; Malioutov and Barzilay, 2006; Purver et al., 2006; Purver, 2011).

Functional structure and *automated functional segmentation* aims to identify sections within a discourse that serve different functions. These functions are genre-specific. In the case of scientific journals, high-level sections generally include the *Background* (work that motivates the objectives of the work and/or the hypothesis or claim being tested), followed by its *Methods*, and *Results*, ending with a *Discussion* of the results or outcomes, along with conclusions to be drawn. Finergrained segments might include the the advantage of a new method (*method-new-advantage*) or of an old method (*method-old-advantage*) or the disadvantage of one or the other (Liakata et al., 2010). Again, segmentation is usually done on a sentence-by-sentence basis, with sentences not assumed to fill more than one function. Methods for functional segmentation have employed specific cue words and phrases, as well as more general language models (Burstein et al., 2003; Chung, 2009; Guo et al., 2010; Kim et al., 2010; Lin et al., 2006; McKnight and Srinivasan, 2003; Ruch et al., 2007; Mizuta et al., 2006; Palau and Moens, 2009; Teufel and Moens, 2002; Teufel et al., 2009; Agarwal and Yu, 2009). The BIO approach to sequential classication (Beginning/Inside/Outside) used in Named Entity Recognition has also proved useful (Hirohata et al., 2008), recognizing that the way the start of a functional segement is signalled may differ from how it is continued.

Note that topic segmentation and functional segmentation are still not always distinguished. For example, in (Jurafsky and Martin, 2009), the term *discourse segmentation* is used to refer to any segmentation of a discourse into a "high-level" linear structure. Nevertheless, segmentation by function exploits different features (and in some cases, different methods) than segmentation by topic, so they are worth keeping distinct.

Attention to event structure and the identification of events within a text is a more recent phenomena, after a hiatus of over twenty years. Here we just point to work by (Bex and Verheij, 2010; Chambers and Jurafsky, 2008; Do et al., 2011; Finlayson, 2009).

The *automated identification of discourse relations* aims to identify discourse relations such as CONDITION and MOTIVATION, as in Example 4, and CONTRAST and SUCCESSION, as in Example 5. These have also been called *coherence relations* or *rhetorical relations*. Methods used depend on whether or not a text is taken to be divisible into a covering sequence of a non-overlapping *discourse units* related to adjacent units by *discourse relations* as in Rhetorical Structure Theory (Mann and Thompson, 1988) or to both adjacent and non-adjacent units as in the Discourse GraphBank (Wolf and Gibson, 2005). If such a cover is assumed, methods involve parsing a text into units using lexical and punctuational cues, followed by labelling the relation holding between them (Marcu, 2000a; Marcu, 2000b; Wolf and Gibson, 2005). If text is not assumed to be divisible into discourse units, then methods involve finding evidence for discourse relations (including both explicit words and phrases, and clausal and sentential adjacency) and their arguments, and then labelling the sense of the identified relation (Elwell and Baldridge, 2008; Ghosh et al., 2011; Lin et al., 2010; Lin, 2012; Prasad et al., 2010a; Wellner, 2008; Wellner and Pustejovsky, 2007).

3.2 Resources for discourse structure

All automated systems for segmenting and labelling text are grounded in data — whether the data has informed the manual creation of rules or has been a source of features for an approach based on machine learning. In the case of topic structure and high-level functional structure, there is now a substantial amount of data that is freely available. For other types of discourse structure, manual annotation has been required and, depending on the type of structure, different amounts are currently available.

More specifically, work on topic structure and segmentation has been able to take advantage of the

large, free, still-growing *wikipedia*, where articles on similar topics tend to show similar explicit segmentation into sub-topics. This is certainly the case with the English wikipedia. If similar wikipedia evolving in other languages lack explicit segmentation, it may be that cross-lingual techniques may be able to project explicit segmentation from English-language articles.

With respect to high-level functional structure, some work on automated segmentation has been able to exploit explicit author-provided indicators of structure, such as the author-structured abstracts now required by bio-medical journals indexed by MedLine. Researchers have used these explicitly structured abstracts to segment abstracts that lack explicit structure (Chung, 2009; Guo et al., 2010; Hirohata et al., 2008; Lin et al., 2006).

For all other kinds of discourse structures, dedicated manual annotation has been required, both for segmentation and labelling, and many of these resources have been made available for other researchers. For *fine-grained functional structure*, there is the ART corpus (Liakata et al., 2010)[2].

For *discourse relations* annotated in the RST framework, there is the RST Discourse TreeBank of English text (Carlson et al., 2003), available through the Linguistic Data Consortium (LDC), as well as similarly annotated corpora in Spanish (da Cunha et al., 2011), Portugese (Pardo et al., 2008) and German (Stede, 2004).

For *discourse relations* annotated in the *lexically-grounded* approach first described in (Webber and Joshi, 1998), there is the Penn Discourse TreeBank (Prasad et al., 2008) in English, as well as corpora in Modern Standard Arabic (Al-Saif and Markert, 2010; Al-Saif and Markert, 2011), Chinese (Xue, 2005; Zhou and Xue, 2012), Czech (Mladová et al., 2008), Danish (Buch-Kromann et al., 2009; Buch-Kromann and Korzen, 2010), Dutch (van der Vliet et al., 2011), Hindi (Oza et al., 2009), and Turkish (Zeyrek and Webber, 2008; Zeyrek et al., 2009; Zeyrek et al., 2010). Also available are discourse-annotated journal articles in biomedicine (Prasad et al., 2011) and discourse-annotated dialogue (Tonelli et al., 2010)).

---

[2]http://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

4 New challenges

Although the largely empirically-grounded, multi-structure view of discourse addresses some of the problems that previous computational approaches encountered, it also reveals new ones, while leaving some earlier problems still unaddressed.

4.1 Evidence for discourse structures

The first issue has to do with what should be taken as evidence for a particular discourse structure. While one could simply consider all features that can be computed reliably and just identify the most accurate predictors, this is both expensive and, in the end, unsatisfying.

With *topic structure*, content words do seem to provide compelling evidence for segmentation, either using language models or semantic relatedness. On the other hand, this might be improved through further evidence in the form of *entity chains*, as explored earlier in (Kan et al., 1998), but using today's more accurate approaches to automated coreference recognition (Strube, 2007; Charniak and Elsner, 2009; Ng, 2010).

Whatever the genre, evidence for *function structure* seems to come from the frequency and distribution of closed-class words, particular phrases (or phrase patterns), and in the case of speech, intonation. So, for example, Niekrasz (2012) shows that what he calls *participant-relational features* that indicate the participants relationships to the text provide convincing evidence for segmenting oral narrative by the type of narrative activity taking place. These features include the distribution and frequency of first and second person pronouns, tense, and intonation. But much work remains to be done in this area, in establishing what provides reliable evidence within a genre and what evidence might be stable across genres.

Evidence for discourse relations is what we have given significant thought to, as the Penn Discourse TreeBank (Prasad et al., 2008) and related corpora mentioned in Section 3.2 aim to ground each instance of a discourse relation in the evidence that supports it. The issue of evidence is especially important because none of these corpora has yet been completely annotated with discourse relations. Completing the annotation and developing robust

automated segmentation techniques requires identifying what elements of the language provide evidence for coherence relations, and under what conditions.

The two main types of evidence for discourse relations in English are the presence of a discourse connective and sentence adjacency. Discourse connectives annotated in the PDTB 2.0 come from a list of subordinating and coordinating conjunctions, and discourse adverbials — a subset of those identified by Forbes-Riley *et al* (2006). Subsequently, Prasad *et al.* (2010b) used Callison-Burch's technique for identifying syntax-constrained paraphrases (Callison-Burch, 2008) to identify additional discourse connectives, some of which don't appear in the PDTB corpus and some of which appear in the corpus but were not identified and annotated as discourse connectives. English isn't alone in lacking a complete list of discourse connectives: While German has the massive *Handbuch de deutschen Konnektoren* (Pasch et al., 2003), even this resource has been found to be incomplete through clever application of automated tagging and word-alignment of parallel corpora (Versley, 2010).

Evidence for discourse relations in the PDTB also comes from lexical or phrasal elements that are outside the initial set of conjunctions and discourse adverbials. This evidence has been called *alternative lexicalization* or *AltLex* (Prasad et al., 2010b), and includes (in English) clause-initial *what's more* (Example 6) and *that means* (Example 7).

(6) A search party soon found the unscathed aircraft in a forest clearing much too small to have allowed a conventional landing. *What's more*, the seven mail personnel aboard were missing. [wsj_0550]

(7) The two companies each produce market pulp, containerboard and white paper. *That means* goods could be manufactured closer to customers, saving shipping costs, he said. [wsj_0317]

The discovery of these other forms of evidence[3] raises the question of **when** it is that a word or phrase signals a discourse relation. For example, only 15 of the 33 tokens of *that means* in the PDTB were annotated as evidence of a discourse relation. While the

---

[3]which English is not alone in having, cf. (Rysova, 2012)

three paragraph-initial instances were left unannotated due to resource limitations (ie, no paragraph initial sentences were annotated unless they contained an explicit discourse connective), the majority were ignored because they followed an explicit connective.

As Wiebe's example (5) showed, there can be multiple explicit discourse connectives in a clause, each of which is evidence for a separate discourse relation (albeit possibly between the same arguments). All of these are annotated in the PDTB – eg, both *but* and *then* in

(8) Congress would have 20 days to reject the package with a 50% majority, but then a President could veto that rejection. [wsj_1698]

The question is whether an *AltLex* in the context of an explicit connective also provides evidence of a distinct discourse relation — for example, with the conjunction with *But* in

(9) At a yearling sale, a buyer can go solo and get a horse for a few thousand dollars. But *that means* paying the horse's maintenance; on average, it costs $25,000 a year to raise a horse. [wsj_1174]

As noted above, the PDTB 2.0 also admits sentence adjacency as evidence for one, or even two, implicit discourse relations, as in

(10) *And some investors fault Mr. Spiegel's life style*; [Implicit = because, for instance] **he earns millions of dollars a year and flies around in Columbia's jet planes**. [wsj_0179]

Here, the implicit token of *because* is associated with a discourse relation labelled CONTINGENCY.CAUSE.REASON, while the implicit token of *for instance* is associated with one labelled EXPANSION.RESTATEMENT.SPECIFICATION.

The question is whether sentence adjacency could also serve as evidence for a distinct discourse relation, even when there is also an explicit discourse adverbial, as in the following three instances of *instead*. Here, Ex. 11 can be paraphrased as *And instead*, Ex. 12 as *But instead*, and Ex.13 as *So instead*.

(11) But many banks are turning away from strict price competition. Instead, they are trying to

build customer loyalty by bundling their services into packages and targeting them to small segments of the population. [wsj_0085]

(12) The tension was evident on Wednesday evening during Mr. Nixon's final banquet toast, normally an opportunity for reciting platitudes about eternal friendship. <u>Instead</u>, Mr. Nixon reminded his host, Chinese President Yang Shangkun, that Americans haven't forgiven China's leaders for the military assault of June 3-4 that killed hundreds, and perhaps thousands, of demonstrators. [wsj_0093]

(13) Since stars are considerably more massive than planets, such wobbles are small and hard to see directly. <u>Instead</u>, Dr Marcy and others like him look for changes that the wobbles cause in the wavelength of the light from the star. [*The Economist*, 10 November 2007]

These examples suggest that the presence of an explicit connective should not, in all cases, be considered evidence for the absense of an implicit connective. Once the set of explicit connectives have been identified that can co-occur with each other (including *for example* and *for instance*, as well as *instead*), automated parsers for coherence relations can be made to consider the presence of an implicit connective whenever one of these is seen.

4.2 Variability in discourse annotation

Another issue relates to variability in annotating discourse structure: Inter-annotator agreement can be very low in annotating pragmatic and discourse-related phenomena. While we will illustrate the point here in terms of annotating coherence relations, for other examples, the general point is illustrated by papers given at the DGfS Workshop on *Beyond Semantics*[4]. and in an upcoming special issue of the journal *Discourse and Dialogue* devoted to the same topic.

The Penn *Wall Street Journal* corpus contains twenty-four (24) reports of *errata* in previously-appearing articles. Twenty-three (23) consist of a single pair of sentences, with no explicit discourse connective signalling the relation between them.[5]

---

[4]http://www.linguistics.ruhr-uni-bochum.de/beyondsem/

[5]The other report contains three sentences, again with no explicit connectives.

One sentence reports the error, and the other, the correct statement – e.g.

(14) VIACOM Inc.'s loss narrowed to $21.7 million in the third quarter from $56.9 million a year ago. Thursday's edition misstated the narrowing. [wsj_1747]

In twenty of the errata (class **C1**), the correct statement is given in the first sentence and the error, in the second; In the other three (class **C2**), it is the other way around. One might think that the two sentences in the twenty **C1** reports would be annotated as having the same discourse relation holding between them, and the same with the two sentences in the three **C2** reports. But that is not the case: The twenty **C1** reports presented to annotators at different times ended up being labelled with **six** different discourse relations. There was even variability in labelling the three members of the **C2** class: They were labelled with one discourse relation, and one with a completely different one.

What should one conclude from this variability? One possibility is that there is one right answer, and annotators just vary in their ability to identify it. This would mean it would be beneficial to have a large troop of annotators (so that the majority view could prevail). Another possibility is that there is more than one right answer, which would imply multi-label classification so that multiple labels could hold to different degrees. A third possibility reflects the view from *Beyond Semantics* that it is often very hard to transfer results from theoretical linguistics based on toy examples to naturally-occurring texts. In this case, variability is a consequence of the still *exploratory nature* of much discourse annotation. In the case of errata, while clearly **some** relation holds between the pair of sentences, it may actually not be any of those used in annotating the PDTB. That is, as Grosz and Sidner (1986b) argued several years ago, the sentences may only be related by their *communicative intentions* – one sentence intended to draw the reader's attention to the specific error that was made (so that the reader knows what was mis-stated), the other intended to correct it. One might then take the sense annotation of discourse relations as still exploratory in the wide range of corpora being annotated with this information (cf. Section 3.2).

4.3 Systematic relations between discourse structures

Fortunately for approaches to automated discourse structure recognition, the lack of isomorphism between different discourse structures does not necessarily mean that they are completely independent. This belief that different aspects of discourse would be related, is what led Grosz and Sidner (1986b) to propose a theory that linked what they called the *intentional structure* of discourse, with its linguistic structure and with the reader or listener's cognitive attentional structure.

With respect to the different types of discourse structure considered here, (Lin, 2012) has considered the possibility of systematic relations between Teufel's *Argumentative Zone* labelling of scientific texts in a corpus developed for her PhD thesis (Teufel, 1999) and PDTB-style discourse relations, both within and across sentences. This is certainly worth additional study, for the value it can bring to automated methods of discourse structure recognition.

4.4 Intentional structure

When computational discourse processing turned to machine learning methods based on reliably-identifiable features, it abandoned (at least temporarily) the centrality of pragmatics and speaker intentions to discourse. That is, there were few or no features that directly indicated or could serve as reliable proxies for what role speaker intended his/her utterance to play in the larger discourse. But both Niekrasz' work on meeting segmentation (Section 4.1 and the discussion in Section 4.2 of errata and variability in their labelling draws new attention to this old question, and not just to Moore and Pollack's observation (Section 3) that *intentional* and *informational* characterizations may confer different, non-isomorphic structures over a text. It may also be the case that neither structure may provide a complete cover: A new visit is warranted.

4.5 Discourse and inference

Not only were intentions abandoned in the move to data-intensive methods, so was inference and issues of how readers and listeners recover information that isn't explicit. What's missing can be an unmentioned event, with classic examples coming from the *restaurant script* (Lehnert, 1977), where someone enters a restaurant, sits down at a table and gives their order to a waiter, where unmentioned *inter alia* is an event in which the person becomes informed of what items the restaurant has to offer, say through being given a menu. Or it can be an unmentioned fact, such as that *program trading* involves the computer-executed trading of a basket of fifteen or more stocks. The latter explains the annotation of an implicit EX-PANSION.RESTATEMENT.GENERALIZATION relation between the two sentences in

(15) "You're not going to stop the idea of trading a basket of stocks," says Vanderbilt's Prof. Stoll. "Program trading is here to stay, and computers are here to stay, and we just need to understand it." [wsj_0118]

The problem here with inference is when labelling an implicit coherence relation requires inferred information about its arguments, those arguments may have quite different features than when all the information needed to label the relation is explicit.

5 Conclusion

There are still large challenges ahead for computational discourse modelling. But we are hopeful that greater openness to how information is conveyed through discourse, as well as richer modelling techniques developed for other problems, will allow needed progress to be made. If we can improve system performance in recognizing the roles that utterances are meant to play in discourse in one genre, perhaps it will help us generalize and transport this intention recognition between genres. We also hope for progress in finding more ways to take advantage of unannotated data in discourse research; in understanding more about inter-dependencies between features of different types of discourse structure; in continuing to carry out related computational discourse research and development in multiple languages and genres, so as to widen the access to the knowledge gained; and in exploiting discourse in Language Technology applications, including information extraction and SMT.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *Proceedings, 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Amal Al-Saif and Katja Markert. 2011. Modelling discourse relations for Arabic. In *Proceedings, Empirical Methods in Natural Language Processing*, pages 736–747.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge UK.

Yves Bestgen. 2006. Improving text segmentation using Latent Semantic Analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

Floris Bex and Bart Verheij. 2010. Story schemes for argumentation about the facts of a crime. In *Proceedings, AAAI Fall Symposium on Computational Narratives*, Menlo Park CA. AAAI Press.

Susan E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting, Association for Computational Linguistics*, pages 155–162, Stanford University, Stanford CA.

Matthias Buch-Kromann and Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 127–131, July.

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

Jill Burstein and Martin Chodorow. 2010. Progress and new directions in technology for automated essay evaluation. In R Kaplan, editor, *The Oxford Handbook of Applied Linguistics*, pages 487–497. Oxford University Press, 2 edition.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 18:32–39.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt & R. Smith, editor, *Current Directions in Discourse and Dialogue*. Kluwer, New York.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings, Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 789–797.

Eugene Charniak and Micha Elsner. 2009. Em works for pronoun anaphora resolution. In *Proc. European Chapter of the Association for Computational Linguistics*.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David Karger. 2009. Global models of document structure using latent permutations. In *Proceedings, Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 371–379.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for text segmentation. In *EMNLP '01: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 109–117.

Grace Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 10(9), February.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the rst spanish treebank. In *Proc. 5th Linguistic Annotation Workshop*, pages 1–10, Portland OR.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings, Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

James Eales, Robert Stevens, and David Robertson. 2008. Full-text mining: Linking practice, protocols and articles in biological research. In *Proceedings of the BioLink SIG, ISMB 2008*.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identication with connective specic rankers. In *Proceedings of the IEEE Conference on Semantic Computing (ICSC-08)*.

Mark Finlayson. 2009. Deriving narrative morphologies via analogical story merging. In *Proceedings, 2nd International Conference on Analogy*, pages 127–136.

Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23:55–106.

George Foster, Pierre Isabelle, and Roland Kuhn. 2010. Translating structured documents. In *Proceedings of AMTA*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*.

Sucheta Ghosh, Sara Tonelli, Giuseppe Riccardi, and Richard Johansson. 2011. End-to-end discourse parser evaluation. In *Proceedings, IEEE Conference on Semantic Computing (ICSC-11)*.

Barbara Grosz and Candace Sidner. 1986a. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara Grosz and Candace Sidner. 1986b. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara Grosz and Candace Sidner. 1990. Plans for discourse. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1986. Towards a computational theory of discourse interpretation. Widely circulated unpublished manuscript.

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

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts. In *Proceedings of the 2010 BioNLP Workshop*, July.

Marti Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Kenji Hirohata, Naoki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientic abstracts using conditional random fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 381–388.

W. Lewis Johnson and Jeff Rickel. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *Int'l J. Artificial Intelligence in Education*, 11:47–78.

Dan Jurafsky and James Martin. 2009. *Speech and Language Processing*. Prentice-Hall, Englewood Cliffs NJ, 2 edition.

Min-Yen Kan, Judith Klavans, and Kathleen McKeown. 1998. Linear segmentation and segment significance. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Andrew Kehler. 1997. Current theories of centering for pronoun intepretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475.

Rodger Kibble and Richard Power. 2000. An integrated framework for text planning and pronominalisation. In *Proc. of the First International Conference on Natural Language Generation*, pages 77–84, Mitzpe Ramon, Israel, June.

Su Nam Kim, David Martinez, and Lawrence Cavedon. 2010. Automatic classification of sentences for evidence based medicine. In *Proc. ACM 4th Int'l Workshop on Data and Text Mining in Biomedical informatics*, pages 13–22.

Alistair Knott, Jon Oberlander, Mick O'Donnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T Sanders, J Schilperoord, and W Spooren, editors, *Text Representation:Linguistic and psycholinguistic aspects*, pages 181–196. John Benjamins Publishing.

Wendy Lehnert. 1977. A conceptual theory of question answering. In *Proc 5th International Joint Conference on Artificial Intelligence*, pages 158–164.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL Workshop on BioNLP*, pages 65–72.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical report, Department of Computing, National University of Singapore, November. http://arxiv.org/abs/1011.0835.

Ziheng Lin. 2012. *Discourse Parsing: Inferring Discourse Structure, Modelling Coherence, and its Applications*. Ph.D. thesis, National University of Singapore.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*.

Jean Mandler. 1984. *Stories, scripts, and scenes: Aspects of schema theory*. Lawrence Erlbaum Associates, Hillsdale NJ.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000a. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Daniel Marcu. 2000b. *The theory and practice of discourse parsing and summarization*. MIT Press.

Mstislav Maslennikov and Tat-Seng Chua. 2007. A multi-resolution framework for information extraction from free text. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 592–599. Association for Computational Linguistics.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the AMIA Annual Symposium*, pages 440–444.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75:468487.

Lucie Mladová, Šárka Zikánová, and Eva Hajičová. 2008. From sentence to discourse: Building an annotation scheme for discourse based on the Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Johanna Moore and Martha Pollack. 1992. A problem for RST: The need for multi-level discouse analysis. *Computational Linguistics*, 18(4):537–544.

Johanna Moore. 1995. *Participating in Explanatory Dialogues*. MIT Press, Cambridge MA.

Megan Moser and Johanna Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first 15 years. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden.

John Niekrasz. 2012. *Toward Summarization of Communicative Activities in Spoken Conversation*. Ph.D. thesis, University of Edinburgh.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In *Proc. 3rd ACL Language Annotation Workshop (LAW III)*, Singapore, August.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proc. 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107. ACM.

Martha Palmer, Carl Weir, Rebecca Passonneau, and Tim Finin. 1993. The kernal text understanding system. *Artificial Intelligence*, 63:17–68.

Thiago Alexandre Salgueiro Pardo, Maria das Gracas Volpe Nunes, and Lucia Helena Machado Rino. 2008. Dizer: An automatic discourse analyzer for brazilian portuguese. *Lecture Notes in Artificial Intelligence*, 3171:224–234.

Renate Pasch, Ursula Brausse, Eva Breindl, and Ulrich Wassner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proc., 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):300–363.

Livia Polanyi and Martin H. van den Berg. 1996. Discourse structure and discourse interpretation. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 113–131, University of Amsterdam.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and et al. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010a. Exploiting scope for shallow discourse parsing. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010b. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings, International Conf. on Computational Linguistics (COLING)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The Biomedical Discourse

Relation Bank. *BMC Bioinformatics*, 12(188):18 pages. http://www.biomedcentral.com/1471-2015/12/188.

Matthew Purver, Tom Griffiths, K.P. Körding, and Joshua Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings, International Conf. on Computational Linguistics (COLING) and the Annual Meeting of the Association for Computational Linguistics*, pages 17–24.

Matthew Purver. 2011. Topic segmentation. In Gokhan Tur and Renato de Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley, Hoboken NJ.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, and et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2–3):195–200.

David Rumelhart. 1975. Notes on a schema for stories. In Dan Bobrow and Alan Collins, editors, *Representation and Understanding: Studies in Cognitive Science*. Academic Press, New York.

Magdalena Rysova. 2012. Alternative lexicalizations of discourse connectives in czech. In *Proc. 8th Int'l Conf. Language Resources and Evaluation (LREC 2012)*.

Roger Schank and Robert Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale NJ.

Penni Sibun. 1992. Generating text without trees. *Computational Intelligence*, 8(1):102–122.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain, July.

Manfred Stede. 2008. RST revisted: Disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *'Subordination' versus 'Coordination' in Sentence and Text*, pages 33–59. John Benjamins, Amsterdam.

Manfred Stede. 2012. *Discourse Processing*. Morgan & Claypool Publishers.

Michael Strube. 2007. Corpus-based and machine learning approaches to anaphora resolution. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference*, pages 207–222. John Benjamins Publishing.

Joel Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings, Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Gian Lorenzo Thione, Martin van den Berg, Livia Polanyi, and Chris Culy. 2004. Hybrid text summarization: combining external relevance measures with structural analysis. In *Proceedings of the ACL 2004 Workshop Text Summarization Branches Out*.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proc. 7th Int'l Conf. Language Resources and Evaluation (LREC 2010)*.

Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. In *Bochumer Linguistische Arbeitsberichte*, pages 157–171.

Steven Vere and Timothy Bickmore. 1990. A basic agent. *Computational Intelligence*, 6(1):41–60.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC)*. NODALIDA.

Marilyn Walker, Aravind Joshi, and Ellen Prince. 1997. *Centering in Discourse*. Oxford University Press, Oxford, England.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In *Coling/ACL Workshop on Discourse Relations and Discourse Markers*, pages 86–92, Montreal, Canada.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*.

Ben Wellner. 2008. *Sequence Models and Ranking Methods for Discourse Parsing*. Ph.D. thesis, Brandeis University.

Janyce Wiebe. 1993. Issues in linguistic segmentation. In *Workshop on Intentionality and Structure in Discourse Relations, Association for Computational Linguistics*, pages 148–151, Ohio StateUniversity.

Terry Winograd. 1973. A procedural model of language understanding. In Roger Schank and Ken Colby, editors, *Computer Models of Thought and Language*,

pages 152–186. W.H. Freeman. Reprinted in Grosz et al. (eds), Readings in Natural Language Processing. Los Altos CA: Morgan Kaufmann Publishers, 1986, pp.249-266.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31:249–287.

William Woods. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the AFIPS National Computer Conference*, pages 457–471, Montvale NJ. AFIPS Press.

William Woods. 1978. Semantics and quantification in natural language question answering. In *Advances in Computers*, volume 17, pages 1–87. Academic Press, New York.

Nianwen Xue. 2005. Annotating discourse connectives in the chinese treebank. In *ACL Workshop on Frontiers in Corpus Annotation II*, Ann Arbor MI.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR6)*.

Deniz Zeyrek, Ümut Deniz Turan, Cem Bozsahin, Ruket Çakıcı, and et al. 2009. Annotating Subordinators in the Turkish Discourse Bank. In *Proceedings of the 3rd Linguistic Annotation Workshop (LAW III)*.

Deniz Zeyrek, Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, İhsan Yalçınkaya, and Ümut Deniz Turan. 2010. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW III)*.

Yuping Zhou and Nianwen Xue. 2012. Pdtb-style discourse annotation of chinese text. In *Proc. 50th Annual Meeting of the ACL*, Jeju Island, Korea.