Natural Language Engineering **18** (4): 437–490. © Cambridge University Press 2011
doi:10.1017/S1351324911000337

Discourse structure and language technology

B. WEBBER[1], M. EGG[2] and V. KORDONI[3]

[1]*School of Informatics*
University of Edinburgh, Edinburgh, UK
e-mail: bonnie@inf.ed.ac.uk
[2]*Department of English and American Studies*
Humboldt University, Berlin, Germany
e-mail: markus.egg@anglistik.hu-berlin.de
[3]*German Research Centre for Artificial Intelligence (DFKI GmbH) and*
Department of Computational Linguistics, Saarland University, Saarbrcken, Germany
e-mail: kordoni@coli.uni-saarland.de

(*Received 8 November 2010; revised 24 October 2011; accepted 26 October 2011; first published online 8 December 2011*)

Abstract

An increasing number of researchers and practitioners in Natural Language Engineering face the prospect of having to work with entire texts, rather than individual sentences. While it is clear that text must have useful structure, its nature may be less clear, making it more difficult to exploit in applications. This survey of work on discourse structure thus provides a primer on the bases of which discourse is structured along with some of their formal properties. It then lays out the current state-of-the-art with respect to algorithms for recognizing these different structures, and how these algorithms are currently being used in Language Technology applications. After identifying resources that should prove useful in improving algorithm performance across a range of languages, we conclude by speculating on future discourse structure-enabled technology.

1 Introduction

Given that language carries information in its structures – morphological, phonological, syntactic, etc., it is fitting that Language Technology (LT) can exploit these structures in two ways. On the one hand, LT can operate on the units provided by structure – for example, the syllable structure of unknown words in text-to-speech (TTS), or the named entities used in relation extraction. On the other hand, it can use structure as a guide to the location of useful information – for example, using dependency structures or parse trees to find the fillers of particular semantic roles.

Less exploited are structures that span multiple sentences – that is, *discourse structures* and *dialogue structures* – but this too is changing. This survey focuses on the former. Early work on *discourse structure* lacked both the huge amounts of text that have become available electronically as a source of empirical evidence and the data-intensive methods and shared resources (e.g., annotated corpora) that allow this evidence to be found and used. With growing amounts of data, methods

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

and resources, discourse structure can now be understood in ways amenable to computational treatment. Thus, the goal of the current survey is to

(1) introduce the reader to discourse structures, their constituent elements, and those of their formal properties relevant to LT (Section 2);
(2) characterize the state-of-the-art algorithms that use discourse structures to find, extract, and/or interpret information in multi-sentence texts (Section 3);
(3) describe current applications of these algorithms in LT tasks (Section 4);
(4) indicate resources under development that should support enhanced algorithmic performance across a range of languages and genres (Section 5);
(5) speculate on future developments (Section 6).

2 Framework

In this section, we present the framework we use in the rest of the paper, specifically addressing the basic questions of what discourse is (Section 2.1), what discourse structures are and what they structure (Section 2.2), and what formal properties of discourse structures are relevant to LT (Section 2.3).

2.1 What is discourse?

Discourse commonly comprises a sequence of sentences, although it can be found even within a single sentence – for example, the connected sequence of *eventualities* (*states* and *events*) described in the following:

(1) If they're drunk and meant to be on parade and you go to their room and they're lying in a pool of piss, then you lock them up for a day.
(*The Independent*, 17 June 1997)

Within a discourse, the patterns formed by its sentences mean that the whole conveys more than the sum of its separate parts. While each sentence in (2) is a simple assertion

(2) Don't worry about the world coming to an end today. It is already tomorrow in Australia (Charles Schulz),

the latter is meant as the REASON for not worrying.

Another point about discourse is that it exploits language features, which allow speakers to specify that they are

- talking about something they have talked about before in the same discourse;
- indicating a relation that holds between the states, events, beliefs, etc. presented in the discourse;
- changing to a new topic or resuming one from earlier in the discourse.

Features that allow speakers to reveal that they are talking about something already under discussion include *anaphoric expressions*, such as the pronouns in

(3) The police are not here to create disorder. **They** are here to preserve **it**. (attributed to Yogi Berra),

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

and the *ellipsis* in

(4) Pope John XXIII was asked 'How many people work in the Vatican?' He is said to have replied, 'About **half**'.

Language features that allow a speaker to specify a relation that holds between the states, events, beliefs, etc. presented in the discourse include *subordinating conjunctions* such as 'until' and *discourse adverbials* such as 'as a result', as in the following:

(5) Men have a tragic genetic flaw. **As a result**, they cannot see dirt **until** there is enough of it to support agriculture.
(paraphrasing Dave Barry, *The Miami Herald*, 23 November 2003)

Finally, language features in discourse that allow speakers to specify a change to a new topic or resumption of an earlier one include what are called *cue phrases* or *boundary features*, as in

(6) Man is now able to fly through the air like a bird
He's able to swim beneath the sea like a fish
He's able to burrow beneath the ground like a mole
Now if only he could walk the Earth like a man,
This would be paradise.
(Lyrics to *This would be Paradise*, Auf der Maur)

While these same features of language can appear in an isolated sentence, such as

(7) When Pope John XXIII was asked 'How many people work in the Vatican?', **he** is said to have replied, 'About **half**',

the sentence usually has multiple clauses, across which these language features serve the same functions.

Given these properties, it is reasonable to associate discourse

(1) with a sequence of sentences,
(2) which conveys more than its individual sentences through their relationships to one another, and
(3) which exploits special features of language that enable discourse to be more easily understood.

Work on **discourse structure** focuses primarily on the second of these.

2.2 What is structured in discourse structures?

Discourse structures are the *patterns* that one sees in multi-sentence (multi-clausal) texts. Recognizing these pattern(s) in terms of the elements that compose them is essential to correctly deriving and interpreting information in the text.

The elements may be *topics*, each about a set of entities and what is being said about them. Or they may be *functions*, each realized by one or more clauses, where

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

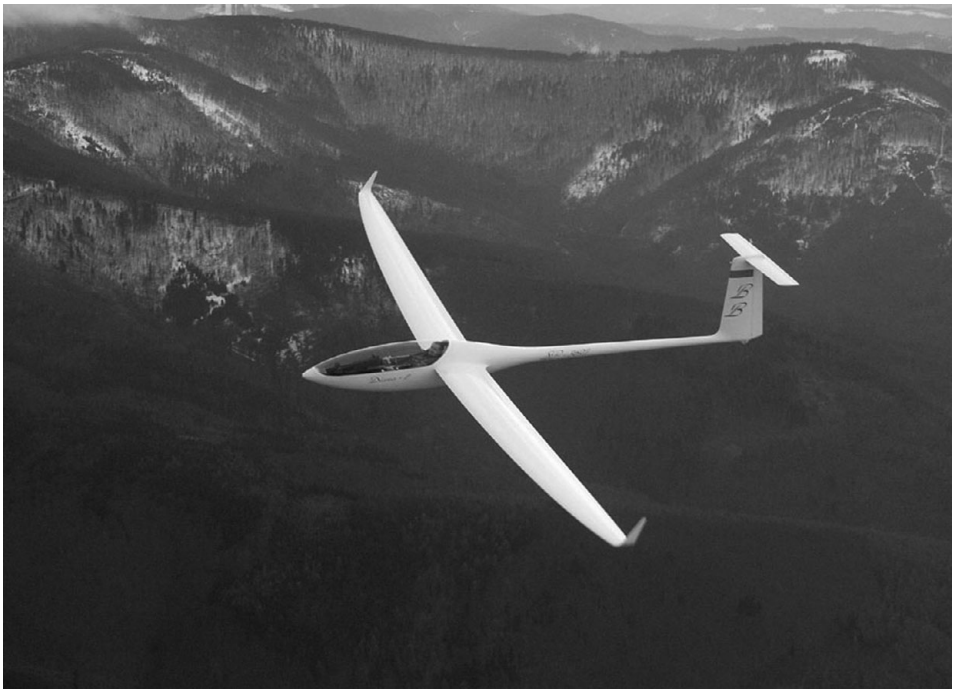


Fig. 1. A glider aloft (http://www.idn.org.pl/users/pawinski/szybowiec_Diana.jpg).

the function served may be with respect to the discourse as a whole or some other segment of discourse. Or they may be *events*, *states*, and/or other *eventualities* and their spatio-temporal relations. Feeding into these are not only clauses and sentences but low-level structures above the sentence, variously called *coherence relations*, *discourse relations*, or *rhetorical relations*. Here we briefly describe each of these structures in turn.

2.2.1 Topics

Discourse can be structured by its *topics*, each comprising a set of entities and a limited range of things being said about them. Topic structure is common in the *expository text* found in text books and encyclopedias. A topic can be characterized by the question it addresses, as in the following text about gliders (Figure 1) from http://simple.wikipedia.org/wiki/Glider, here annotated with question-phrased topics:

(8) Gliders are aircraft which do not have a motor. They are sometimes called 'sailplanes'. ⇒ **What defines a glider**
Gliders are controlled by their pilots by using control-sticks. Some gliders can only carry one person, but some gliders can carry two persons. In gliders with two seats, each pilot has a control-stick. Gliders always have seats for the pilots. ⇒ **How gliders are controlled**

Gliders have long wings so that they will only lose height slowly. In some places the air goes up faster than the glider is going down. The pilot of a

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

glider can make it climb by flying to these places. Good pilots can travel long distances by always finding rising air. Some pilots race each other over hundreds of kilometres each day. Other pilots just fly for fun. ⇒ **How gliders travel**

Gliders cannot get into the air by themselves. They are pulled into the air by an aircraft with a motor or they are pulled up by motor on the ground.
⇒ **How gliders get off the ground**

Each topic involves a set of entities, which may (but do not have to) change from topic to topic. Here the entities consist of gliders; then gliders, their pilots, and passengers; then gliders and their means of propulsion; and then gliders and their launch mechanisms. This aspect of structure has been modelled as *entity chains* (Barzilay and Lapata 2008) – each a sequence of expressions that refer to the same entity. There are several entity chains in the text about gliders:

$$\begin{gathered} \textit{Gliders} \rightarrow \textit{They} \rightarrow \textit{Gliders} \rightarrow \textit{their} \rightarrow \ldots \\ \textit{some places} \rightarrow \textit{these places} \\ \textit{their pilots} \rightarrow \textit{each pilot} \rightarrow \textit{the pilots} \end{gathered}$$

Where a set of entity chains ends and another set starts has been used as evidence that the discourse has moved from one topically oriented segment to another (Kan, Klavans and McKeown 1998). Patterns of entity chains can also be characteristic of particular types of discourse, and therefore be of value in assessing the quality of automatically generated text (Section 4.5).

Low-level evidence for the topic structure of discourse comes from the strong correlation between topic and lexical usage, which Halliday and Hasan (1976) call *lexical cohesion*. This can involve word repetition, the use of semantically related words, such as *hypernyms* (more general terms), *hyponyms* (more specific terms), *synonyms* (terms with a similar sense), and *meronyms* (terms that refer to a part of a given whole), as well as the use of words that bear more general associations such as between *fly* and *air*, or *aircraft* and *pilot*. Lexical cohesion can simply be a matter of the *density* of related terms within a segment, or of particular patterns of related terms, such as *lexical chains* (Barzilay and Elhadad 1997; Galley *et al.* 2003; Clarke and Lapata 2010), defined as sequences of semantically related words – e.g.,

$$\begin{gathered} \textit{Gliders} \xrightarrow{\textit{hypernymy}} \textit{aircraft} \\ \textit{Gliders} \xrightarrow{\textit{synonymy}} \textit{sailplanes} \\ \textit{Gliders} \xrightarrow{\textit{meronymy}} \textit{wings} \end{gathered}$$

Over time, topic structures can become *conventionalized* in their coverage and order, as shown in the sub-headings of Wikipedia articles about US states (Figure 2). Having such a conventional order simplifies both the problem of finding particular types of information and that of assembling information into a natural text, as discussed in Section 3.1. In Natural Language Generation, conventionalized topic structures were used in early work on text generation (McKeown 1985; Paris 1988).

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

	Wisconsin	Louisiana	Vermont
1	Etymology	Etymology	Geography
2	History	Geography	History
3	Geography	History	Demographics
4	Demographics	Demographics	Economy
5	Law and government	Economy	Transportation
6	Economy	Law and government	Media
7	Municipalities	Education	Utilities
8	Education	Sports	Law and government
9	Culture	Culture	Public health
10	...	...	...

Fig. 2. Sub-headings of Wikipedia articles about US states.

2.2.2 *Functions*

Discourse can also be structured by the *functions* served by its elements – that is, by their role(s) in the communication. One can identify very general communicative roles or intentions – e.g., assist another element in fulfilling its communicative role(s) – or more specific ones – e.g., present a conclusion drawn from previous discourse elements. There are also communicative roles that are specific to particular genres – e.g., *identify the alleged offenses* in textual descriptions of criminal cases (Moens, Uyttendaele and Dumortier 1999), and *present advantage of methodology* in scientific research papers (Liakata *et al.* 2010). In fact, one aspect of many textual *genres* is their *conventionalized high-level functional structure*.

The genre of news reports is a good example of this: The functions served in their conventionalized *inverted pyramid structure* comprise (1) a summary of who is involved, what took place, when and where it took place, why it happened, and (optionally) how, all found in an initial *lede paragraph*; (2) a sequence of paragraphs that provide more details about these points in the *body*; and (3) less important information presented in a final *tail*.

Another example is the structure in which eventualities and their spatio-temporal relations are presented in text – conventionally, in their chronological order. Departures from this convention produce very striking effects – as in films, such as *Memento* and *Betrayal* where scenes appear in reverse chronological order.

A third example that has been a significant focus of attention in LT is scientific research papers – and more recently, their abstracts,[1] the high-level functions served in their sequence of sections comprise (1) the background for the research, which motivates its objectives and/or the hypothesis being tested (*Background*); (2) the methods or study design used in the research (*Methods*); (3) the results or outcomes (*Results*); and (4) a discussion of the results and a presentation of conclusions to be drawn (*Discussion*). Algorithms for segmenting the functional structure of genre-specific texts are discussed in Section 3.1.

[1] http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

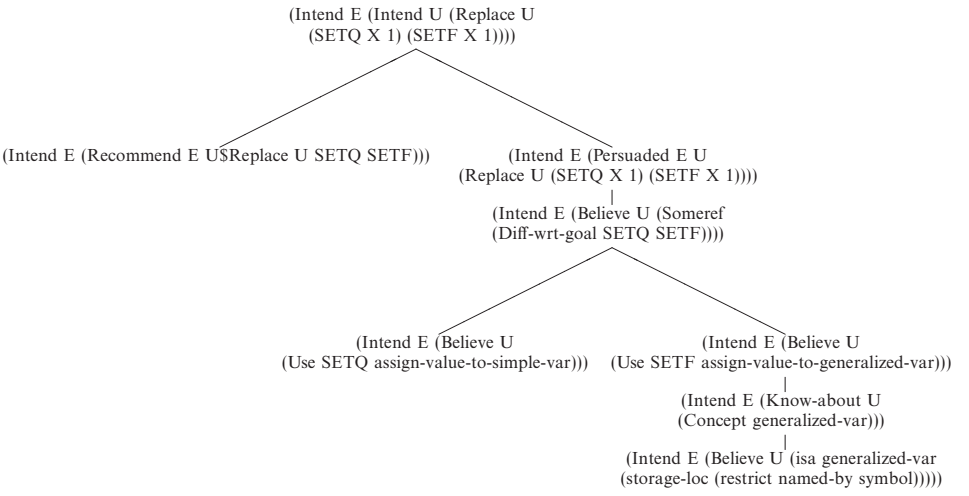


Fig. 3. Intentional structure of Example 9.

The least conventionalized functional structure is a wide-open reflection of the speaker's communicative intentions and their relations to each another. This produces a more complex *intentional structure*, which was a major focus of important work in the 1980s and 1990s (Grosz and Sidner 1986, 1990; Moore and Paris 1993; Moser and Moore 1996; Lochbaum 1998). The kind of tree structure commonly assumed for intentional structure (Section 2.3.1) is illustrated in Figure 3. Moore and Paris (1993) give this as the structure underlying the utterance in Example 9 made by someone tutoring a student in the programming language LISP.

(9) You should replace (SETQ X 1) with (SETF X 1). SETQ can only be used to assign a value to a simple-variable. SETF can be used to assign a value to any generalized-variable. A generalized-variable is a storage location that can be named by any access function (Moore and Paris 1993).

Since the recognition of *intentional structure* seems to require extensive modelling of human intentions and their relations, there has been little empirical work on this area of functional structure except in the context of dialogue (Section 6.1).

2.2.3 Eventualities

Discourse can be structured by *eventualities* (descriptions of *events* and *states*) and their spatio-temporal relations. One can find such structure in news reports, in the *Methods* section of a scientific paper, in *accident reports*, and, more generally, in *Narrative*, as should be clear from its definition:

'A perceived sequence of nonrandomly connected events, i.e., of described states or conditions which undergo change (into some different states or conditions).' (Toolan 2006)

As with the previous two structuring devices (i.e., *topics* and *functions*), patterns of eventualities may be conventionalized, as in Propp's analysis of Russian folk tales (Propp 1968) in terms of common morphological elements such as

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

- *an interdiction is addressed to the protagonist*, where the hero is told not to do something;
- *the interdiction is violated*, where the hero does it anyway;
- *the hero leaves home*, on a search or journey;
- *the hero is tested or attacked*, which prepares the way for receiving a magic agent or helper.

Or they may be more open, associated with individual psychologies or environmental factors. In the 1970s and the early 1980s, there was considerable interest in the linear and hierarchical structuring inherent in narrative, expressed in terms of *story schemata* (Rumelhart 1975), *scripts* (Schank and Abelson 1977) and *story grammars* (Kintsch and van Dijk 1978; Mandler 1984). This was motivated in part by the desire to answer questions about stories – in particular, allowing a reader to 'fill in the gaps', recognizing events that had to have happened, even though they have not been explicitly mentioned in the narrative.

Because of the clear need for extensive world knowledge about events and their relations, and (as with *intentional structure*, Section 2.2.2) for extensive modelling of human intentions and their relations, there has been little empirical work in this area until very recently (Chambers and Jurafsky 2008; Finlayson 2009; Bex and Verheij 2010; Do, Chan and Roth 2011).

2.2.4 Discourse relations

Discourse also has low-level structure corresponding to *discourse relations* that hold either between the semantic content of two units of discourse (each consisting of one or more clauses or sentences) or between the speech act expressed in one unit and the semantic content of another.[2] This semantic content is an *abstract object* (Asher 1993) – a proposition, a fact, an event, a situation, etc. Discourse relations can be explicitly signalled through explicit discourse connectives as in

(10) The kite was created in China, about 2,800 years ago. **Later** it spread into other Asian countries, like India, Japan and Korea. **However**, the kite only appeared in Europe by about the year 1600. (http://simple.wikipedia.org/wiki/Kite)

Here the adverbial 'later' expresses a SUCCESSION relation between the event of creating kites and that of kites spreading to other Asian countries, while the adverbial 'however' expresses a CONTRAST relation between the spread of kites into other Asian countries and their spread into Europe. One can consider each of these a *higher order* predicate-argument structure, with the discourse connective ('later' and 'however') conveying the predicate with two abstract objects expressing its arguments.[3]

[2] The smallest unit of discourse, sometimes called a *basic discourse unit* (Polanyi *et al.* 2004b) or *elementary discourse unit* or EDU (Carlson, Marcu and Okurowski 2003), usually corresponds to a clause or nominalization, or an anaphoric or deictic expression referring to either, but other forms may serve as well – cf. Section 5.1.6.

[3] No discourse connective has yet been identified in any language that has other than two arguments.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Relations can also be signalled implicitly through utterance adjacency, as in

(11) Clouds are heavy. The water in a cloud can have a mass of several million tons. (http://simple.wikipedia.org/wiki/Cloud)

Here the second utterance can be taken to either ELABORATE or INSTANTIATE the claim made in the adjacent first utterance. (In terms of their *intentional structure*, the second utterance can be taken to JUSTIFY the first.) Algorithms for recovering the structure associated with discourse relations are discussed in Section 3.2, and its use in text summarization and sentiment analysis is discussed in Sections 4.1 and 4.4, respectively.

2.3 Properties of discourse structure relevant to LT

The structures associated with *topics*, *functions*, *eventualities*, and *discourse relations* have different formal properties that have consequences for automatically extracting and encoding information. The ones we discuss here are complexity (Section 2.3.1), coverage (Section 2.3.2), and symmetry (Section 2.3.3).

2.3.1 Complexity

Complexity relates to the challenge of recovering structure through segmentation, chunking, and/or parsing (Section 3). The earliest work on discourse structure for both text understanding (Kintsch and van Dijk 1978; Grosz and Sidner 1986; Mann and Thompson 1988; Grosz and Sidner 1990) and text generation (McKeown 1985; Dale 1992; Moore 1995; Walker *et al.* 2007) viewed it as having a *tree structure*. For example, the natural-sounding recipes are automatically generated in Dale (1992), such as

(12) *Butter Bean Soup*
Soak, drain and rinse the butter beans. Peel and chop the onion. Peel and chop the potato. Scrape and chop the carrots. Slice the celery. Melt the butter. Add the vegetables. Saute them. Add the butter beans, the stock and the milk. Simmer. Liquidise the soup. Stir in the cream. Add the seasonings. Reheat,

have a structure isomorphic to a hierarchical plan for producing them (Figure 4), modulo *aggregation* of similar daughter nodes that can be realized as a single conjoined unit (e.g., 'Soak, drain and rinse the butter beans').

At issue among advocates for tree structure underlying all (and not just some) types of discourse was what its nodes corresponded to. In the Rhetorical Structure Theory (RST) (Mann and Thompson 1988), terminal nodes projected to *elementary discourse units* (cf. Footnote 2), while a non-terminal corresponded to a complex discourse unit with particular *rhetorical relations* holding between its daughters. The original version of RST allowed several relations to simultaneously link different discourse units into a single complex unit. More recent applications of RST – viz., Carlson *et al.* (2003) and Stede (2004) – assume only a single relation linking the immediate constituents of a complex unit, which allows the identification of non-terminal nodes with discourse relations.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

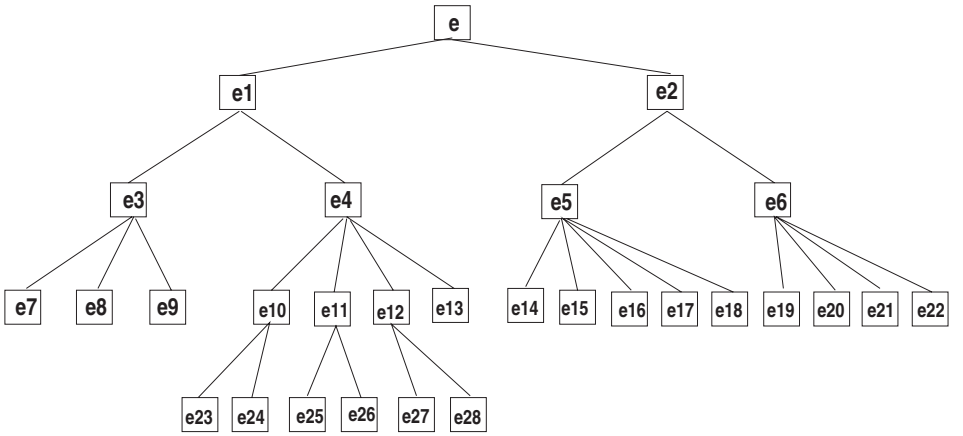


Fig. 4. Discourse structure of recipe for *butter bean soup* from Dale (1992).

In Dale (1992), each node in the tree (both non-terminal and terminal) corresponded to the next step in a plan to accomplish its parent. In *text grammar* (Kintsch and van Dijk 1978), as in sentence-level grammar, higher level non-terminal constituents (each with a communicative goal) rewrite as a sequence of lower level non-terminals with their own communicative goals. And in Grosz and Sidner's (1986) work on the *intentional structure* of discourse, all nodes corresponded to *speaker intentions*, with the communicative intention of a daughter node supporting that of its parent, and *precedence* between nodes corresponding to the need to satisfy the earlier intention before one that follows.

Other proposed structures were nearly trees, but with some nodes having multiple parents, producing sub-structures that were *directed acyclic graphs* rather than trees. Other 'almost tree' structures display crossing dependencies. Both are visible among the discourse relations annotated in the Penn Discourse TreeBank (PDTB) (Lee *et al.* 2006, 2008). The most complex discourse structures are the *chain graphs* found in the Discourse GraphBank (Wolf and Gibson 2005). These graphs reflect an annotation procedure in which annotators were allowed to create discourse relations between any two discourse segments in a text without having to document the basis for the linkage.

At the other extreme, topic-oriented texts have been modelled with a simple *linear* topic structure (Sibun 1992; Hearst 1997; Barzilay and Lee 2004; Malioutov and Barzilay 2006). Linear topic structures have also been extended to serve as a model for the descriptions of objects and their historical contexts given on museum tours (Knott *et al.* 2001). Here within a linear sequence of segments that take up and elaborate on a previously mentioned entity are more complex *tree-structured* descriptions as shown in Figure 5.

2.3.2 Coverage

Coverage relates to how much of a discourse belongs to the structural analysis. For example, since every part of a discourse is about something, all of it belongs

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

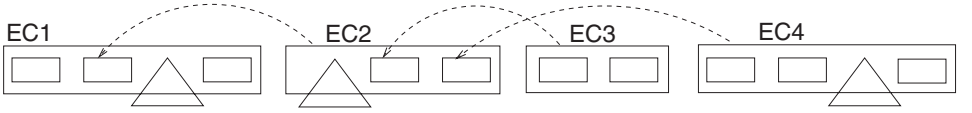


Fig. 5. Illustration of the mixed linear/hierarchical structure presented by Knott *et al.* (2001) for extended descriptions. EC stands for *entity chain*, and the dotted arrows link the *focussed entity* in the next chain with its introduction earlier in the text.

somewhere within a topic segmentation (Section 3.1). So segmentation by topic provides a *full cover* of a text. On the other hand, the structure associated with discourse relations and recovered through discourse chunking (Section 3.2) may only be a *partial cover*. The latter can be seen in the conventions used in annotating the PDTB (Prasad *et al.* 2008):

(1) An attribution phrase is only included in the argument to a discourse relation if the relation holds between the attribution and another argument (e.g., a contrast between what different agents said or between what an agent said and what she did etc.) or if the attribution is conveyed in an adverbial (e.g., 'according to government figures'). Otherwise, it is omitted.
(2) A *Minimality Principle* requires that an argument only includes that which is needed to complete the interpretation of the given discourse relation. Any clauses (e.g., parentheticals, non-restrictive relative clauses, etc.) not so needed are omitted.

This is illustrated in Example 13: Neither the attribution phrase (boxed) nor the non-restrictive relative clause that follows is included in either argument of the discourse relation associated with *But*.

(13) '*I'm sympathetic with workers who feel under the gun*', says Richard Barton of the Direct Marketing Association of America, which is lobbying strenuously against the Edwards beeper bill. 'But **the only way you can find out how your people are doing is by listening**'. (wsj_1058)[4], CatalogEntry=LDC95T7

2.3.3 Symmetry or asymmetry

Symmetry has to do with the importance of different parts of a discourse structure – whether all parts have equal weight. In particular, RST (Mann and Thompson 1988) takes certain discourse relations to be **asymmetric**, with one argument (the *nucleus*) more essential to the purpose of the communication than its other argument (the *satellite*). For example, looking ahead to Section 4.1 and Example 22, whose RST analysis is given in Figure 6, the second clause of the following sentence is taken to be more essential to the communication than its first clause, and hence the sentence is analyzed as *satellite–nucleus*:

[4] Labels of the form wsj_xxxx refer to sections of the *Wall Street Journal Corpus*, http://www.ldc.upenn.edu

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

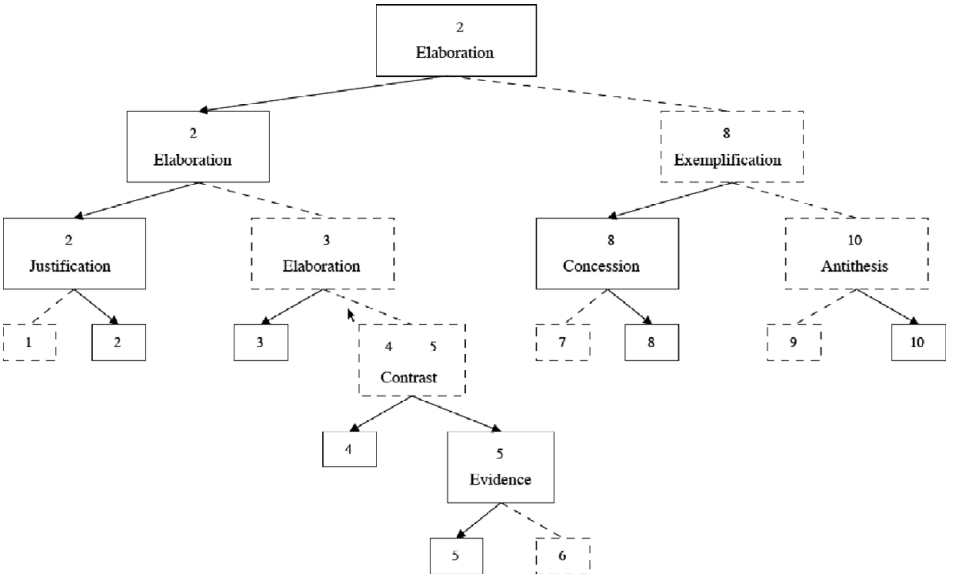


Fig. 6. Discourse structure of Example (22).

(14) Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.

The belief that satellites can thus be removed without harm to the essential content of a text underlies the RST-based approaches to *extractive summarization* (Daume III and Marcu 2002; Uzêda, Pardo and Nunes 2010) as discussed in Section 4.1. However, Section 5.2 presents arguments from Stede (2008b) that RST's concept of nuclearity conflates too many notions that should be considered separately.

3 Algorithms for discourse structure

In this section, we discuss algorithms for recognizing or generating various forms of discourse structures. The different algorithms reflect different properties that are manifested by discourse structure. We start with *discourse segmentation*, which divides a text into a linear sequence of adjacent topically coherent or functionally coherent segments (Section 3.1). Then we discuss *discourse chunking*, which identifies the structures associated with discourse relations (Section 3.2), concluding in Section 3.3 with a discussion of *discourse parsing*, which (like sentence-level parsing) constructs a complete and structured cover over a text.

3.1 Linear discourse segmentation

We discuss segmentation into a linear sequence of topically coherent or functionally coherent segments in the same section in order to highlight similarities and differences in the methods and features that are employed by each segment.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

3.1.1 Topic segmentation

Being able to recognize topic structure was originally seen as benefitting information retrieval (Hearst 1997). More recently, its potential value in segmenting lectures, meetings, or other speech events has come to the fore, making such oral events more amenable to search (Galley *et al.* 2003; Malioutov and Barzilay 2006).

Segmentation into a linear sequence of topically coherent segments generally assumes that the topic of a segment will differ from that of adjacent segments (adjacent spans that share a topic are taken to belong to the same segment.) It is also assumed that topic constrains lexical choice, either of all words of a segment or just its content words (i.e., excluding *stop-words*).

Topic segmentation is based on either *semantic-relatedness*, where words within a segment are taken to relate to each other more than to words outside the segment (Hearst 1997; Choi, Wiemer-Hastings and Moore 2001; Galley *et al.* 2003; Bestgen 2006; Malioutov and Barzilay 2006), or *topic models*, where each segment is taken to be produced by a distinct and compact lexical distribution (Purver *et al.* 2006; Eisenstein and Barzilay 2008; Chen *et al.* 2009). In both approaches, segments are taken to be sequences of sentences or pseudo-sentences (i.e., fixed-length strings), whose relevant elements may be all the words or just the content words.

All *semantic-relatedness approaches* to topic segmentation involve (1) a *metric* for assessing the semantic relatedness of terms within a proposed segment; (2) a *locality* that specifies which units within the text are assessed for semantic relatedness; and (3) a *threshold* for deciding how low relatedness can drop before it signals a shift to another segment.

Hearst's (1994, 1997) work on TextTiling is a clear illustration of this approach. Hearst considers several different relatedness metrics before focussing on simple *cosine similarity*, using a vector representation of fixed-length spans in terms of word stem frequencies (i.e., words from which any inflection has been removed). Cosine similarity is computed solely between adjacent spans, and an empirically determined threshold is used to choose segment boundaries.

Choi *et al.* (2001) and Bestgen (2006) use *Latent Semantic Analysis* (LSA) instead of word-stem frequencies in assessing semantic relatedness, again via cosine similarity of adjacent spans. While LSA may be able to identify more lexical cohesion within a segment (increasing intra-segmental similarity), it may also recognize more lexical cohesion across segments (making segments more difficult to separate).

Galley *et al.* (2003) use *lexical chains* to model lexical cohesion, rather than either word-stem frequencies or LSA-based concept frequencies. Even though their lexical chains exploit only term repetitions, rather than the wider range of relations noted in Section 2.2.1, lexical chains are still argued to better represent topics than simple frequency by virtue of their structure: Chains with more repetitions can be weighted more highly with a bonus for chain *compactness* – of two chains with the same number of repetitions, the shorter can be weighted more highly. Compactness captures the locality of a topic, and the approach produces the state-of-the-art performance. A similar approach is taken by Kan *et al.* (1998), though

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

based on *entity chains*. This enables pronouns to be included as further evidence for intra-segmental semantic similarity.

Rather than considering the *kinds* of evidence used in assessing semantic relatedness, Malioutov and Barzilay (2006) experiment with *locality*: Instead of just considering relatedness of adjacent spans, they consider the relatedness of all spans within some large neighborhood whose size is determined empirically. Rather than making segmentation decision-based simply on the (lack of) relatedness of the next span, they compute a weighted sum of the relatedness of later spans (with weight attenuated by distance) and choose boundaries based on minimizing lost relatedness using *min-cut*. This allows for more gradual changes in topic than do those approaches that only consider the next adjacent span.

The more recent Bayesian *topic modelling approach* to discourse segmentation is illustrated in the work of Eisenstein and Barzilay (2008). Here each segment is taken to be generated by a topically constrained language model over word stems with stop-words removed and with words in a segment modelled as draws from the model. Eisenstein and Barzilay also attempt to improve their segmentation through modelling how *cue words*, such as *okay*, *so*, and *yeah*, are used in speech to signal topic change. For this, a separate draw of a topic-neutral cue phrase (including none) is made at each topic boundary. Their system, though slow, does produce performance gains on both written and spoken texts.

Turning to texts in which the order of topics has become more or less conventionalized (Section 2.2.1), recent work by Chen *et al.* (2009) uses a latent topic model for unsupervised learning of global discourse structure that makes neither the too weak assumption that topics are randomly spread through a document (as in the work mentioned above) nor the too strong assumption that the succession of topics is fixed. The global model they use (the *generalized Mallows model*) biases toward sequences with a similar ordering by modelling a distribution over the space of topic permutations, concentrating probability mass on a small set of similar ones. Unlike in Eisenstein and Barzilay (2008), word distributions are not just connected to topics, but to discourse-level topic structure. Chen *et al.* (2009) show that on a corpus of relatively conventionalized articles from Wikipedia their generalized Mallows model outperforms Eisenstein and Barzilay's (2008) approach.

For an excellent overview and survey of topic segmentation, see Purver (2011).

3.1.2 Functional segmentation

As outlined in Section 2.2.2, functional structure ranges from the conventionalized, high-level structure of particular text genres to the non-formulaic *intentional structure* of a speaker's own communicative intentions. Because of the open-ended knowledge of the world and of human motivation needed for recognizing intentional structure, recent empirical approaches to recognize functional structure have focussed on producing a flat segmentation of a discourse into labelled functional regions. For this task, stop words, linguistic properties like tense, and extra-linguistic features, such as citations and figures, have proved beneficial to achieve good performance.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Most of the computational work on this problem has been done on the genre of *biomedical abstracts*. As we have noted in Section 2.2.2, scientific research papers commonly display explicitly labelled sections that deal (in order) with (1) the background for the research, which motivates its objectives and/or the hypothesis being tested (*Background*); (2) the methods or study design used in the research (*Methods*); (3) the results or outcomes (*Results*); and (4) a discussion thereof, along with conclusions to be drawn (*Discussion*).

Previously such section labels were not found in biomedical abstracts. While these abstracts have been erroneously called *unstructured* (in contrast with the *structured abstracts* whose sections are explicitly labelled), it is assumed that both kinds have roughly the same structure. This means that a corpus of structured abstracts can serve as relatively free *training data* for recognizing the structure inherent in unstructured abstracts.[5]

The earliest of this work (McKnight and Srinivasan 2003) treated functional segmentation of biomedical abstracts as an individual sentence classification problem, usually with a sentence's rough location within the abstract (e.g., start, middle, end) as a feature. Later work took it as a problem of learning a sequential model of sentence classification with sentences related to *Objectives* preceding those related to *Methods*, which in turn precede those related to *Results*, ending with ones presenting *Conclusions*. Performance on the task improved when Hirohata *et al.* (2008) adopted the Beginning/Inside/Outside (*BIO*) model of sequential classification from Named Entity Recognition. The BIO model recognizes that evidence that signals the start of a section may differ significantly from evidence of being inside the section. Although results are not directly comparable since all research reported to date has been trained and tested on different corpora, both Chung (2009) and Hirohata and colleagues (2008) report accuracy that is 5-to-10 points higher than McKnight and Srinivaasan (2003).

Work is also being carried out on automating a fine-grained functional labelling of scientific research papers (Teufel and Moens 2002; Mizuta *et al.* 2006; Liakata *et al.* 2010). This work has shown that high-level functional segmentation is not strongly predictive of all the fine-grained functional labels of sentences within a given segment. (see also Guo *et al.* 2010, who compare high-level functional segmentation of research papers and abstracts with these two fine-grained functional labelling schemes on a hand-labelled corpus of 1,000 abstracts on cancer risk assessment.) On the other hand, attention to larger patterns of fine-grained functional labels could be a first step toward reconstructing an *intentional structure* of what the writer is trying to achieve.

Progress is also being made on recovering and labelling the parts of texts with other conventional functional structures – legal arguments consisting of sentences expressing premises and conclusions (Palau and Moens 2009), student

[5] Not all structured abstracts use the same set of section labels. However, most researchers (McKnight and Srinivasan 2003; Lin *et al.* 2006; Ruch *et al.* 2007; Hirohata *et al.* 2008; Chung 2009) opt for a set of four labels, usually some variant of *Objectives*, *Methods*, *Results*, and *Conclusions*.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

essays (Burstein, Marcu and Knight 2003), and the full text of biomedical articles (Agarwal and Yu 2009).

3.2 Discourse chunking

By *discourse chunking*, we refer to recognizing units within a discourse such as *discourse relations* that are not assumed to provide a full *cover* of the text (Section 2.3.2). Discourse chunking is thus a lightweight approximation to *discourse parsing*, discussed in Section 3.3.

Underpinning any approach to recognizing discourse relations in a text are answers to three questions:

(1) Given a language, what affixes, words, terms, and/or constructions can signal discourse relations, and which tokens in a given discourse actually do so?
(2) Given a token that signals a discourse relation, what are its arguments?
(3) Given such a token and its arguments, what sense relation(s) hold between the arguments?

Here we address the first two questions. The third will be discussed with discourse parsing (Section 3.3), since the issues are the same. By distinguishing and ordering these questions, we are not implying that they need to be answered separately or in that order in practice: Joint solutions may work even better.

Given a language, what elements (e.g., affixes, words, terms, constructions) can signal discourse relations? Some entire part of speech classes can signal discourse relations, although particular tokens can serve other roles as well. For example, all coordinating and subordinating conjunctions signal discourse relations when they conjoin clauses or sentences, as in

(15) a. Finches eat seeds, and/but/or robins eat worms.
b. Finches eat seeds. But today, I saw them eating grapes.
c. While finches eat seeds, robins eat worms.
d. Robins eat worms, just as finches eat seeds.

With other parts of speech, only a subset may signal discourse relations. For example, with adverbials, only *discourse adverbials* such as 'consequently' and 'for example' signal discourse relations:

(16) a. Robins eat worms and seeds. Consequently they are omnivores.
b. Robins eat worms and seeds. Frequently they eat both simultaneously.

While both pairs of sentences above bear a discourse relation to each other, only in 16(a) is the type of the relation (RESULT) signalled by the adverb. In the ELABORATION relation expressed in 16(b), the adverb just conveys how often the situation holds. Discourse relations can also be signalled by special constructions that Prasad, Joshi and Webber (2010b) call *alternative lexicalizations* – for example,

- *This/that <be> why/when/how* <S> (e.g., 'That's why we're here'.)
- *This/that <be> before/after/while/because/if/etc.* <S> (e.g., 'That was after we arrived'.)

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

- *The reason/result <be>* <S> (e.g., 'The reason is that we want to get home'.)
- *What's more* <S> (e.g., 'What's more, we've taken too much of your time'.)

Identifying all the elements in a language, including *alternative lexicalizations*, that signal discourse relations is still an open problem (Prasad *et al.* 2010b). One option is to use a list of known discourse connectives to automatically find other ones. Prasad *et al.* (2010b) show that additional discourse connectives can be discovered through monolingual paraphrase via back-translation (Callison-Birch 2008). Versley (2010) shows how *annotation projection*[6] can be used to both annotate German connectives in a corpus and discover those not already included in the *Handbuch der Deutschen Konnektoren* (Pasch *et al.* 2003).

Not all tokens of a given type may signal a discourse relation.[7] For example, 'once' only signals a discourse relation when it serves as a subordinating conjunction (Example 17(a)), not as an adverbial (Example 17(b)):

(17) a. Asbestos is harmful *once* it enters the lungs. (*subordinating conjunction*)
b. Asbestos was *once* used in cigarette filters. (*adverb*)

Although tokens may appear ambiguous, Pitler and Nenkova (2009) found that for English, discourse and non-discourse usage can be distinguished with at least 94% accuracy.

Identifying discourse relations also involves recognizing its two arguments.[8] In the PDTB, these two arguments are simply called

- **Arg2**: the argument from text syntactically bound to the connective;
- *Arg1* : the other argument.

Because **Arg2** is defined in part by its syntax, the main difficulty comes from *attribution* phrases, which indicate that the semantic content is 'owned' by some agent. This ownership may or may not be a part of the argument. The attribution phrase in Example 18 (here, boxed) is not a part of **Arg2**, while in Example 19, both *Arg1* and **Arg2** include their attribution phrase.

(18) *We pretty much have a policy of not commenting on rumors*, and I think **that falls in that category**. (wsj_2314)

(19) *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**. (wsj_0098)

Because *Arg1* need not be adjacent to **Arg2**, it can be harder to recognize. Firstly, like pronouns, *anaphoric* discourse adverbials may take as its *Arg1* an entity introduced earlier in the discourse rather than one that is immediately adjacent – for example

[6] In annotation projection, texts in a source language are annotated with information (e.g., POS-tags, coreference chains, semantic roles, etc.), which the translation model then projects in producing the target text. Other uses of annotation projection are mentioned in Section 6.2.

[7] The same is true of discourse markers (Petukhova and Bunt 2009).

[8] As noted in Section 2.2.4, no discourse connective has yet been identified in any language that has other than two arguments.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

(20) On a level site you can provide a cross pitch to the entire slab by *raising one side of the form* (step 5, p. 153), but for a 20-foot-wide drive this results in an awkward 5-inch (20 x 1/4 inch) slant across the drive's width. Instead, **make the drive higher at the center**.

Here, *Arg1* of *instead* comes from just the 'by' phrase in the previous sentence – that is, the drive should be made higher at its center instead of raising one side.

Secondly, annotation of the PDTB followed a *minimality principle* (Section 2.3.2), so arguments need only contain the minimal amount of information needed to complete the interpretation of a discourse relation. In Example 21, neither the quote nor its attribution are needed to complete the interpretation of the relation headed by But, so they can be excluded from *Arg1*. The result is that *Arg1* is not adjacent to **Arg2**.

(21) *Big buyers like Procter & Gamble say there are other spots on the globe and in India, where the seed could be grown.* 'It's not a crop that can't be doubled or tripled', says Mr. Krishnamurthy. But **no one has made a serious effort to transplant the crop**. (wsj_0515)

There is a growing number of approaches to the problem of identifying the arguments to discourse connectives. Wellner (Wellner and Pustejovsky 2007; Wellner 2008) has experimented with several approaches using a 'head-based' dependency representation of discourse that reduces argument identification to simply locating their *heads*.

In one experiment, Wellner (Wellner and Pustejovsky 2007; Wellner 2008) identified discourse connectives and their candidate arguments using a discriminative log-linear ranking model on a range of syntactic, dependency, and lexical features. He then used a log-linear re-ranking model to select the best pair of arguments (*Arg1*–**Arg2**) in order to capture any dependencies between them. Performance on coordinating conjunctions improves through re-ranking from 75.5% to 78.3% accuracy (an 11.4% error reduction), showing the model captures dependencies between *Arg1* and **Arg2**. While performance is significantly worse on discourse adverbials (42.2% accuracy), re-ranking again improves performance to 49% (an 11.8% error reduction). Finally, while performance is the highest on subordinating conjunctions (87.2% accuracy), it is degraded by re-ranking to 86.8% accuracy (a 3% increase in errors). So if dependencies exist between the arguments of subordinating conjunctions, they must be different in kind than those that hold between the arguments to coordinating conjunctions or discourse adverbials.

Wellner (Wellner and Pustejovsky 2007; Wellner 2008) also investigated a fully joint approach to discourse connective and argument identification, which produced a 10%–12% reduction in errors over a model he explored, which identified them sequentially.

Wellner's (Wellner and Pustejovsky 2007; Wellner 2008) results suggest that better performance might come from *connective-specific* models. Elwell and Baldridge (2008) investigate this, using additional features that encode the specific connective (e.g., *but*, *then*, *while*, etc.); the type of connective (coordinating conjunction, subordinating conjunction, discourse adverbial); and local context features, such as

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

the words to the left and right of the candidate and to the left and right of the connective. While Elwell and Baldridge (2008) demonstrate no performance difference on coordinating conjunctions, and slightly worse performance on subordinating conjunctions, performance accuracy improved significantly on discourse adverbials (67.5% vs. 49.0%), showing the value of connective-specific modelling.

More recently, Prasad, Joshi and Webber (2010a) have shown improvements on Elwell and Baldridge's (2008) results by taking into account the **location** of a connective – specifically, improved performance for inter-sentential coordinating conjunctions and discourse adverbials by distinguishing *within-paragraph* tokens from *paragraph-initial* tokens. This is because 4,301/4,373 (98%) of within-paragraph tokens have their *Arg1* in the same paragraph, which significantly reduces the search space. (Paragraphs in the *Wall Street Journal* corpus tend to be very short – an average of 2.17 sentences per paragraph across the 1,902 news reports in the corpus, and an average of three sentences per paragraph across its 104 essays (Webber 2009).) Ghosh *et al.* (2011b) achieve even better performance on recognizing *Arg1* both within and across paragraphs by including **Arg2** labels in their feature set for recognizing *Arg1*.

Although component-wise performance still has a way to go, nevertheless it is still worth verifying that the components can be effectively assembled together. This has now been demonstrated, first in the work of Lin, Ng and Kan (2010), whose end-to-end processor for discourse chunking identifies explicit connectives, their arguments and their senses, as well as implicit relations and their senses (only top eleven sense types, given data sparsity) and attribution phrases, and more recently in the work of Ghosh *et al.* (2011a).

3.3 Discourse parsing

As noted earlier, *discourse parsing* resembles sentence-level parsing in attempting to construct a complete structured cover of a text. As such, only those types of discourse structures that posit more of a cover than a linear segmentation (e.g., RST (Mann and Thompson 1988), Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides 2003), and Polanyi's Theory of discourse structure and coherence (Polanyi *et al.* 2004a)) demand discourse parsing.

Now, any type of parsing requires (1) a way of identifying the basic units of analysis – i.e., *tokenization*; (2) a method for exploring the search space of possible structures and labels for their nodes; and (3) a method for deciding among alternative analyses. Although discourse parsing is rarely described in these terms and tokenization is sometimes taken for granted (as was also true in early work on parsing – cf. Woods (1968)), we hope it nevertheless provides a useful framework for understanding what has been done to date in the area.

3.3.1 Tokenization

Sentence-level parsing of formal written text relies on the fact that sentence boundaries are explicitly signalled, though the signals are often ambiguous. For

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

example, while a period ('full stop') can signal a sentence boundary, it can also appear in abbreviations, decimal numbers, formatted terms, etc. However, this is still less of a problem than that of identifying the units of discourse (*discourse tokenization*) for two reasons:

(1) There is no general agreement as to what constitutes the elementary units of discourse (sometimes called EDUs) or as to what their properties are – e.g., whether or not they admit discontinuities.
(2) Since parsing aims to provide a complete cover for a discourse, when one unit of a discourse is identified as an EDU, what remains in the discourse must also be describable in EDU terms as well. (Note that this is not true in the case of *discourse chunking*, which is not commited to providing a complete cover of a text.)

In the construction of the RST Corpus (Carlson *et al.* 2003), significant attention was given to clearly articulating of rules for tokenizing an English text into EDUs so that they can be applied automatically. In this same RST framework, Sagae (2009) treats discourse tokenization as a binary classification task on each word of a text that has already been parsed into a sequence of dependency structures: The task is to decide whether or not to insert an EDU boundary between the current word and the next. Features used here include, *inter alia*, the current word (along with its POS-tag, dependency label, and direction to its head), and the previous two words (along with their POS-tags and dependency labels). Discourse tokenization by this method resulted in a precision, recall, and F-score of 87.4%, 86%, and 86.7%, respectively, on the testing section of the RST Corpus.

In the work of Polanyi *et al.* (2004b), discourse tokenization is done after sentence-level parsing with the Lexical-Functional Grammar (LFG)-based Xerox Linguistic Environment (XLE). Each sentence is broken up into discourse-relevant units based on lexical, syntactic, and semantic information, and then these units are combined into one or more small discourse trees, called Basic Discourse Unit (BDU) trees, which then play a part in subsequent processing. These discourse units are thus syntactic units that encode a minimum unit of content and discourse function. Minimal functional units include greetings, connectives, discourse markers, and other cue phrases that connect or modify content segments. In this framework, units may be discontinuous or even fragmentary.

Also allowed to be discontinuous are the complex units into which Baldridge, Asher and Hunter (2007) segment discourse – e.g., allowing a complex unit to omit a discourse unit associated with an intervening attribution phrase such as 'officials at the Finance Minstry have said'. However, their experiments on discourse parsing (discussed below) do not treat these complex units as full citizens, using only their first EDU. Tokenization is the manually done gold standard.

Other researchers either assume that discourse segmentation has already been carried out, allowing them to focus on other parts of the process (e.g., Subba, Eugenio and Kim 2006), or they use sentences or clauses as a proxy for basic discourse segments. For example, in order to learn elementary discourse units that should be linked together in a parse tree, Marcu and Echihabi (2002) take as their EDUs

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

two clauses (main and subordinate) associated with unambiguous subordinating conjunctions.

3.3.2 Structure building and labelling

Discourse parsing explores the search space of possible parse structures by identifying how the units of a discourse (elementary and derived) fit together into a structure, with labels usually drawn from some set of semantic and pragmatic sense classes. Structure building and labelling can be done using rules (manually authored or induced through machine learning (e.g., Subba *et al.* 2006), or probabilistic parsing, or even vector-based semantics (Schilder 2002). The process may also exploit preferences, such as a preference for right-branching structures, and/or well-formedness constraints, such as the *right frontier constraint* (Polanyi *et al.* 2004b), which stipulates that the next constituent to be incorporated into an evolving discourse structure can only be linked to a constituent on its right frontier.[9]

In the work of Polanyi *et al.* (2004b), the parser decides where to attach the next BDU into the evolving structure based on a small set of rules that consider syntactic information, lexical cues, structural features of the BDU and the proposed attachment point, and the presence of constituents of incomplete n-ary constructions on the right edge. The approach thus aims to unify sentential syntax with discourse structure so that most of the information needed to assign a structural description to a text becomes available from regular sentential syntactic parsing and regular sentential semantic analysis.

Subba *et al.* (2006) attempt to learn rules for attachment and labelling using Inductive Logic Programming (ILP) on a corpus of manually annotated examples. The resulting rules have the expressive power of first-order logic and can be learned from positive examples alone. Within this ILP framework, labelling (i.e., deciding what discourse relation holds between linked discourse units) is done as a classification task. Verb semantic representations in VerbNet[10] provide the background knowledge needed for ILP and the manually annotated discourse relations between pairs of EDUs serve as its positive examples. Subba and Eugenio (2009) take this a step further, focusing on the genre of instruction manuals in order to restrict the relevant sense labels to a small set. They also demonstrate performance gains through the use of some genre-specific features, including genre-specific verb semantics, suggesting genre-specific discourse parsers as a promising avenue of research.

The covering discourse structures built by Baldridge *et al.* (2007) are nominally based on SDRT (Asher and Lascarides 2003), which allows such structures to be directed graphs with multiply parented nodes and crossing arcs. However, only one of the two discourse parsing experiments described by Baldridge *et al.* (2007) treats

[9] This is actually a simple stack constraint that has previously been invoked in resolving *object anaphora* (Holler and Irmen 2007) and *event anaphora* (Webber 1991), constraining their antecedents to ones in a segment somewhere on the right-hand side of the evolving discourse structure.

[10] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

them as such. Both experiments use *dependency parsing*, which takes each discourse segment as a token whose direct SDRT-labelled dependency on another token must be identified. One experiment uses McDonald's Minimum-Spanning Tree (MST) parser (McDonald, Crammer and Pereira 2005) with a projective algorithm, which cannot recover crossing SDRT dependencies, and the other uses a non-projective algorithm, which can recover crossing SDRT dependencies. Complex units (see the discussion of *tokenization* above) are replaced in both experiments with their first element. Results reported on a manually segmented and labelled corpus of sixty news texts from MUC-6[11] are 23.0%/51.9% labelled/unlabelled accuracy for the non-projective algorithm and 22.8%/50.8% accuracy for the projective algorithm.

For building discourse structures in an RST framework (Mann and Thompson 1988), Marcu (1999) used a decision-tree learner and shallow syntactic features to create a shift-reduce parser for RST structure-building and labelling. Later, Soricut and Marcu (2003) used the output of the Charniak parser as input to a bottom-up chart parsing algorithm for discourse structure. This had a higher accuracy than Marcu's (1999) earlier shift-reduce approach but recovered only *intra-sentential* discourse relations.

Sagae's (2009) approach to structure-building and labelling in the RST framework sends EDUs identified through discourse tokenization (described above) to a transition-based constituent parsing algorithm (Sagae and Lavie 2005). This treats each EDU as an individual token and builds an RST tree using one *shift* and three *reduce* operations. The latter operations require identifying and upwardly percolating *head EDUs*, which requires the prior identification of the nucleus/satellite status of each EDU in the text (cf. Section 2.3.3). As might be expected, Sagae's approach (2009) outperforms both Marcu's (1999) and Soricut and Marcu's (2003) approaches.

Chart parsing is used in the approach to discourse analysis proposed by Schilder (2002). It uses linguistic knowledge based on discourse markers to constrain an underspecified discourse representation. Whatever remains underspecified is then further specified via a vector space model computation of a topicality score for every discourse unit. Sentences in prominent positions, like first sentences of paragraphs etc., are given an adjusted topicality score. Thus, this parsing algorithm assumes that the number of distinct structures that can be constructed over a sequence of n discourse units is exponential in n. Its robustness is served by overt discourse markers. In their absence, other compensating techniques would be required – for instance, an underspecified parse forest of rhetorical trees.

One general issue in work on discourse structure building and labelling is that of having sufficient data to test and/or appropriately weight the features used in attachment and labelling decisions. There have been at least two efforts aimed at using unsupervised machine learning techniques to acquire such data. Marcu and Echihabi (2002) attempted to automatically create useful labelled data from a large unannotated corpus of multi-clause sentences, each sentence containing an unambiguous subordinating conjunction as its sentence-internal discourse connective. They labelled each sentence with the relation signalled unambiguously by its

[11] http://www-nlpir.nist.gov/related projects/muc/

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

connective and assigned it features consisting of all word pairs drawn from the clauses so connected (one from each clause). They then removed the connective from each example and trained a sense recognizer on the now 'unmarked' examples.

Sporleder and Lascarides (2008) extend this approach by adding syntactic features based on POS-tags, argument structure, and lexical features. They report that their richer feature set, combined with a boosting-based algorithm, is more accurate than the original word pairs alone, achieving 57.6% accuracy in a five-way classification task, where Marcu and Echihabi (2002) achieve 49% accuracy in a six-way classification task.

More importantly, Sporleder and Lascarides (2008) consider the validity of a methodology in which artificial 'unmarked' examples are created from ones with explicit unambiguous connectives, and show that it is suspect. Webber (2009) provides further evidence against this methodology, based on significant differences in the distribution of senses across explicit and implicit connectives in the PDTB corpus (e.g., 1,307 explicit connectives expressing CONTINGENCY.CONDITION versus one implicit connective with this sense, and 153 explicit connectives expressing EXPANSION.RESTATEMENT versus 3,148 implicite connectives with this sense). However, the relevant experiment has not yet been done on the accuracy of recognizing unmarked coherence relations based on both Sporleder and Lascarides' (2008) richer feature set and priors for unmarked coherence relations in a corpus like the PDTB.

4 Applications

Here we consider applications of the research presented earlier, concentrating on a few in which discourse structure plays a crucial role – summarization, information extraction (IE), essay analysis and scoring, sentiment analysis, and assessing the naturalness and coherence of automatically generated text. (For a more complete overview of applications of the approach to discourse structure called *Rhetorical Structure Theory*, the reader is referred to Taboada and Mann (2006).)

4.1 Summarization

Document summarization is one of the earliest applications of discourse structure analysis. In fact, much of the research to date on discourse parsing (in both the RST framework and other theories of hierarchical discourse structure) has been motivated by the prospect of applying it to summarization (Ono, Sumita and Miike 1994; Daume III and Marcu 2002). For this reason, we start by describing summarization based on a weighted hierarchical discourse structure (Marcu 2000; Thione *et al.* 2004) and then review other ways in which research on discourse structure has been applied to summarization.

Summarization based on weighted hierarchical discourse structure relies on the notion of *nuclearity* (cf. Section 2.3), which takes one part of a structure, the *nucleus*, to convey information that is more central to the discourse than that conveyed by the rest (one or more *satellites*). As a consequence, a satellite can often be omitted

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

from a discourse without diminishing its readability or altering its content.[12] If a discourse is then taken to be covered by a hierarchical structure of relations, each of which consists of a nucleus and satellites, a partial ordering of discourse elements by importance (or summary worthiness) can then be derived, and a cut-off chosen, above which discourse elements are included in the summary. The length of the summary can thus be chosen freely, which makes summarization *scalable*. Consider the following example from Marcu (2000):

(22) With its distant orbit – 50% farther from the sun than Earth – and slim atmospheric blanket, C_1 Mars experiences frigid weather conditions. C_2 Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees Celsius near the poles. C_3 Only the mid-day sun at tropical latitudes is warm enough to thaw ice on occasion, C_4 but any liquid water formed in this way would evaporate almost instantly C_5 because of the low atmospheric pressure. C_6 Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, C_7 most Martian weather involves blowing dust or carbon dioxide. C_8 Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap. C_9 Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water. C_{10}

The discourse structure tree that Marcu (2000) gives for (22) is depicted in Figure 6. Here the labels of the nuclei in partial trees percolate to their respective mother node. The nucleus of a relation is indicated with a solid line, and the satellite is indicated with a dashed line.

The weight of a discourse segment is then calculated with respect to the labels assigned to tree nodes. Each branching level constitutes an equivalence class of equally important nodes (excepting those that already show up in higher branching levels). The equivalence classes are calculated top-down. For Figure 6, the equivalence classes and their ordering is $2 > 8 > 3,\ 10 > 1, 4, 5, 7,\ 9 > 6$. Consequently, a two-segment summary of (22) should consist of C_2 and C_8, which would be augmented by C_3 and C_{10} in a four-segment summary. While different methods have been suggested in the literature to calculate these weights, Uzêda *et al.* (2010) show that these methods yield similar results.

Approaches to summarization that exploit *configuration* – i.e., the position of discourse segments in the discourse structure and the status of segments as nucleus or satellite can be found in both Marcu (2000) system and the PALSUMM system of Thione *et al.* (2004).

Recently, information on the *discourse relations* that link specific segments was used to distinguish material that should or should not be included in summaries. Louis, Joshi and Nenkova (2010) compare the predictive power of configurational

[12] Note that this does not hold good for all discourse relations, e.g., omitting the premise of a CONDITION relation would severely change a discourse.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

properties of discourse structure against relevant discourse relations for the summary worthiness of specific discourse segments. They conclude that information on discourse configuration is a good indicator for which segments should show up in a summary, whereas discourse relations turn out useful for the identification of material that should be omitted from the summary. Louis and Nenkova (2011) use the discourse relations INSTANTIATION and RESTATEMENT as defined and annotated in the PDTB to identify more general sentences in a text, which they claim are typical for handcrafted but not for automatically generated summaries and hence should be preserved in summaries.

This approach instantiates a set of design choices for approaches to summarization on the basis of discourse structure. First, it is an instance of *extractive* summarization, which selects the most important sentences for a summary. This contrasts with *sentence compression*, which shortens the individual sentences (Mani 2001).

A second design choice involves the goal of the summary: Daume III and Marcu (2002) attempt to derive *informative* summaries that represent the textual content of documents. An alternative goal, useful in summarizing scientific articles, involves highlighting the contribution of an article and relating it to previous work (Teufel and Moens 2002). With *indicative summaries*, the goal is to facilitate the selection of documents that are worth reading (Barzilay and Elhadad 1997).

A third design choice involves assumptions about the document to be summarized. While Daume III and Marcu (2002) assume a *hierarchical* structure, other approaches just take it to be flat (cf. Section 2.2.2). For example, in summarizing scientific papers, Teufel and Moens (2002) assume that a paper is divided into research goal (*aim*), outline of the paper (*textual*), presentation of the paper's contribution (methods, results, and discussion – labelled here *own*), and presentation of other work (*other*). They classify individual sentences for membership in these classes by discourse segmentation (Section 3.1). This strategy is especially fruitful if the summarization concentrates on specific core parts of a document rather than on the document as a whole.

Teufel and Moens (2002) do not assume that all sentences within a given section of the paper belong to the same class (cf. Section 3.1), but they do find that adherence to a given ordering differs by scientific field: Articles in the natural sciences appear more sequential in this respect than the Computational Linguistics articles that they are targeting.

A fourth design decision involves the *type of document* to be summarized. Most summarization work targets either news or scientific articles. This choice has wide ramifications for a summarizer because the structure of these documents is radically different: The 'inverted pyramid' structure of news articles (cf. Section 2.2.2) means that their first sentences are often good summaries, while for scientific articles, core sentences are more evenly distributed. This difference shows, for instance, in the evaluation of Marcu's (2000) summarizer, which was developed on the basis of essays and argumentative text: Its F-score on summarizing scientific articles was up to 9.1 points higher than its F-score on summarizing newspaper articles.

A final design decision involves the way a summarizer *identifies the discourse structure* on which their summarization is based. While Marcu (2000) crucially relies

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

```
                    Name: %MURDERED%
                  Event Type: MURDER
                TriggerWord: murdered
Activating Conditions: passive-verb
  Slots: VICTIM <subject>(human)
    PERPETRATOR<prep-phrase, by>(human)
   INSTRUMENT<prep-phrase, with>(weapon)
```

Fig. 7. Template for extraction of information on murders.

on cue phrases (especially discourse markers) and punctuation for the identification of elementary and larger discourse units, Teufel and Moens (2002) characterize discourse elements by features like location in the document, length, and lexical and phrasal cue elements (e.g., *along the lines of*), and citations.

A third method involves the use of *lexical chains* (Section 2.2.1). Lexical chains can be used for both extraction and compression: For Barzilay and Elhadad (1997), important sentences comprise the first representative element of a strong lexical chain, and it is these sentences that are selected for the summary. For Clarke and Lapata (2010), sentence compression requires that terms from strong chains must be retained. However, there are different ways of calculating the strength of lexical chains. Barzilay and Elhadad (1997) base it on length and homogeneity of the chain, Clarke and Lapata (2010) base it on the amount of sentences spanned over.

The use of lexical chains allows *topicality* to be taken into account to heighten the quality of summaries. Clarke and Lapata (2010) require the entity that serves as the *center* of a sentence (in the sense of the Centering Theory, cf. Section 4.5) be retained in a summary based on sentence compression. Schilder (2002) shows that discourse segments with low topicality (measured in terms of their similarity to the title or a lead text) should occupy a low position in a hierarchical discourse structure that can be used for extractive summarization.

4.2 Information extraction

The task of information extraction is to extract from text-*named entities*[13] relations that hold between them, and event structures in which they play a role. IE systems focus on specific domains (e.g., terrorist incidents) or specific types of relations (e.g., people and their dates of birth, protein–protein interactions). Event structures are often described by templates in IE, where the named entities to be extracted fill in specific slots, as in Figure 7.

Discourse structure can be used to guide the selection of parts of a document which are relevant to IE. This strategy is a part of a larger tendency toward a

[13] Named entities comprise persons, locations, and organizations, but also various numeric expressions, e.g., times or monetary values. The challenge for NLP is to establish the identity of these entites across widely varying ways of referring to them.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

two-step IE, which first identifies relevant regions for a specific piece of information and then tries to extract this piece of information from these regions.

Decoupling these two steps boosts the overall performance of IE systems (Patwardhan and Riloff 2007). Restricting the search for information to relevant parts of a document reduces the number of false hits (which often occur in irrelevant parts) and, consequently, of erroneous multiple retrievals of potential fillers for the same slot. For example, the IE task of finding new results in biomedical articles has the problem that not all the results referred to in a paper are new. Instead, they may be background or findings reported in other papers (Mizuta *et al.* 2006). At the same time, limiting search to only the relevant parts of a text increases confidence because potential candidates are more likely to be correct. This method was also promoted by the insight that in order to extract all the desired information from a scientific article, only the full article suffices (and not only the abstract, as in earlier attempts to do IE for scientific articles).

Much of this work does not consider discourse structure. For example, approaches like Gu and Cercone (2006) or Tamames and de Lorenzo (2010) classify individual sentences for their likelihood of containing extraction relevant material. But discourse structure information has proven to be valuable for this classification if the structure of the documents is strongly *conventionalized*, as for example in scientific articles or legal texts (Section 2.2.2).

Different kinds of discourse structures can be used for IE purposes. Mizuta *et al.* (2006) use a flat discourse structure based on the discourse zoning of Teufel and Moens (2002) for IE from biology articles. While Moens *et al.* (1999) assume that their legal texts have a hierarchical discourse structure that can be described in terms of a *text grammar* (Kintsch and van Dijk 1978), their work on IE from legal texts only use its sequential upper level. In contrast, Maslennikov and Chua's (2007) IE approach uses a hierarchical discourse structure.

More specifically, Mizuta *et al.*'s (2006) goal is to identify the novel contribution of a paper. They note that this cannot be done by merely looking at the section labelled *Results*, as this would produce both false positives and false negatives (Section 3.1.2). Therefore, they adopt discourse zoning (Teufel and Moens 2002) as an initial process to distinguish parts of papers that present previous work from those parts that introduce novel contributions, taking advantage of the fact that zoning results in the attribution of results to different sources (present or previous work). Mizuta *et al.* (2006) then classify the novel contributions of a paper (the *own* class of Teufel and Moens 2002) into subclasses such as *Method*, *Result*, *Insight*, and *Implication*. They conclude by investigating the distribution of these subclasses across the common division of scientific articles into four parts here called *Introduction*, *Materials and Methods*, *Results*, and *Discussion* (cf. Section 3.1).

For this, they hand-annotated twenty biological research papers and correlated the fine-grained subclasses with the four zones. Some of their results confirm expectations, while others do not. For example, while the *Materials and Methods* section consists almost exclusively of descriptions of the author's methods, and more than 90% of these novel methodological contributions are located there, only 50% of results are actually expressed in the *Results* section.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Eales, Stevens and Robertson's (2008) work complements the results of Mizuta *et al.* (2006): Their goal is the extraction of information on protocols of molecular phylogenetic research from biological articles. These protocols describe the methods used in a scientific experiment, they are extracted in order to assess their quality (and along with it, the quality of the entire article).

To this end, Eales *et al.* (2008) rely on two insights: first, the fact that scientific articles typically consist of the four parts mentioned above (*Introduction*, *Materials and Methods*, *Results*, and *Discussion*), and second, the high correlation between the second of these sections and methodological innovations of a scientific paper (as shown by Mizuta *et al.* 2006). They trained a classifier to identify these four sections (discourse zoning), and then concentrated on the *Methods* section to extract information about the way in which a specific piece of research was conducted. They got extremely high precision (97%) but low recall (55%) because the classifier failed to recognize all parts of the documents that belong to the *Methods* section.

A similar kind of exploitation of genre-specific structural conventions can be found in the SALOMON system of Moens *et al.* (1999), which extracts relevant information from criminal cases (with the eventual goal of producing short indicative summaries). The system makes use of the fact that these cases have a highly conventionalized functional structure in which for example victim and perpetrator are identified in text segments preceding the one in which the alleged offences and the opinion of the court are detailed.

The relevant information to be extracted from the documents are the respective offences and their evaluation by the court. It is fairly straightforward to extract this information after the document has been segmented, as the functional label of a segment is strongly predictive of the information it will contain.

Maslennikov and Chua's (2007) approach is different as it assumes a fully hierarchical discourse structure. Their goal is to extract semantic relations between entities, for instance, '*x* is located in *y*'. They point out that extracting these relations on the basis of correlations between these relations and paths through a syntactic tree structure (between the nodes for the constituents that denote these entities) is highly unreliable once these syntactic paths get too long. This is bound to happen once one wants to advance to syntactic units above the clause level.

Therefore, they complement these paths by analogous paths through a hierarchical discourse tree in the RST framework which are derived by Soricut and Marcu's (2003) discourse parser *Spade*. These paths link the elementary discourse units of which the constituents denoting the entities are part. This discourse information is used to filter the wide range of potentially available syntactic paths for linguistic expressions above the clause level (only 2% of which are eventually useful as indicators of semantic relations).

Maslennikov and Chua (2007) show that their inclusion of information from discourse structure leads to an improvement of the F-score from 3% to 7% in comparison to other state-of-the-art IE systems that do not take into account discourse structure. However, this strategy basically amounts to reintroducing clause structure into their system because the EDU structures are typically clausal. Hence,

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

they do not make use of the full discourse hierarchy but restrict themselves to the lower levels of the hierarchy within the confines of individual sentences.

4.3 Essay analysis and scoring

Another application for research on discourse structure is essay analysis and scoring, with the goal of improving the quality of essays by providing relevant feedback. This kind of evaluation and feedback is focussed on the *organizational structure* of an essay, which is a crucial feature of quality. For this application, specific *discourse elements* in an essay must first be identified. These discourse elements are part of a non-hierarchical genre-specific conventional discourse structure (Section 2.2.2). For their identification, probabilistic classifiers are trained on annotated data and evaluated against an unseen part of the data.

A first step is the automatic identification of *thesis statements* (Burstein *et al.* 2001). Thesis statements explicitly identify the purpose of the essay or preview its main ideas. Assessing the argumentation of the essay centers around the thesis statement.

The features used by Burstein *et al.* (2001) to identify thesis statements are their position in the essay, characteristic lexical items, and RST-based properties obtained from discourse parsing (Soricut and Marcu 2003), including for each sentence the discourse relation for which it is an argument and its nuclearity. Their Bayesian classifier could identify thesis statements on unseen data with a precision of .55 and a recall of .46 and was shown to be applicable to different essay topics.

Burstein *et al.* (2003) extend this approach to the automatic identification of all essential discourse elements of an argumentative essay, in particular introductory material, thesis, main point (the latter two making up the thesis statement), supporting ideas, and conclusion. Example 23 illustrates the segmentation into discourse elements of an essay's initial paragraph.

(23)

<**Introductory material**> In Korea, where I grew up, many parents seem to push their children into being doctors, lawyers, engineer etc. </**Introductory material**> <**Main point**>Parents believe that their kids should become what they believe is right for them, but most kids have their own choice and often doesn't choose the same career as their parent's. </**Main point**> <**Support**> I've seen a doctor who wasn't happy at all with her job because she thought that becoming doctor is what she should do. That person later had to switch her job to what she really wanted to do since she was a little girl, which was teaching. </**Support**>

Burstein *et al.* (2003) trained three automated discourse analyzers on this data. The first, a decision-tree analyzer, reused features from Burstein *et al.* (2001) plus explicit lexical and syntactic discourse cues (e.g., discourse markers or syntactic subordination) for the identification of discourse elements. The other two were probabilistic analyzers that associated each essay with the most probable sequence of discourse elements. For example, a sequence with a conclusion at the beginning would have a low probability.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

All three analyzers significantly outperform a naive baseline that identifies discourse elements by position. Even better results were obtained by combining the best analyzers through voting. Performace nevertheless varied by the type of discourse element: For example, for INTRODUCTORY MATERIAL, baseline precision/recall/F-score of 35/23/28 improved to 68/50/57 through voting, while for the CONCLUSION, precision/recall/F-score went from a higher baseline of 56/67/61 to 84/84/84 through voting.

The next step in this thread of research is then to assess the internal coherence of an essay on the basis of having identified its discourse elements. Higgins *et al.* (2004) define coherence in terms of three dimensions of *relatedness* measured as the number or density of terms in the same semantic domain: (1) The individual sentences of the essay must be related to the (independently given) essay question or topic, in particular, those sentences that make up thesis statement, background, and conclusion; (2) specific sentences must be related to each other, e.g., background and conclusion sentences to sentences in the thesis; and (3) the sentences within a single discourse element (e.g., background) should be related to each other.

Higgins *et al.* (2004) use a support vector machine to assess coherence. Evaluated on manually annotated gold-standard data, they found that it is very good on the first dimension when there was high relatedness of sentences to the essay question, with an F-score of .82, and on the second dimension, with an F-score of .84. It was less good at detecting low relatedness of sentences to the essay question (F-score of .51) and low relatedness between sentences (F-score of .34). Further work is needed to assess relatedness along the third dimension.

4.4 Sentiment analysis and opinion mining

Finally, we comment on the roles that we believe discourse structure can play in the increasingly popular areas of *sentiment analysis* and *opinion mining*, including (1) assessing the overall opinion expressed in a review (Turney 2002; Pang and Lee 2005); (2) extracting fine-grained opinions about individual features of an item; and (3) summarizing the opinions expressed in multiple texts about the same item. We believe much more is possible than has been described to date in the published literature.

The simplest use we have come across to date was suggested by Polanyi and Zaenen (2004), and involves taking into account discourse connectives when assessing the positive or negative contribution of a clause. They note, for example, that a positive clause such as 'Boris is brilliant at math' should be considered neutralized in a concession relation such as in

(24) Although Boris is brilliant at math, he is a horrible teacher.

Another simple use we have noticed reflects the tendency for reviews to end with an overall evaluative judgment based on the opinions expressed earlier. Voll and Taboada (2007) have used this to fine-tune their approach to sentiment analysis to give more weight to evaluative expressions at the end of text, reporting approximately 65% accuracy. One can also refine this further by employing an approach like *Appraisal Analysis* (Martin 2000), which distinguishes different dimensions along

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

which opinion may vary, each of which can be assigned a separate score. In the case of *appraisal analysis*, these dimensions are affect (emotional dimension), judgement (ethical dimension), and appreciation (aesthetic dimension).

However, approaches that ignore discourse structure will encounter problems in cases like Example (25), which express a positive verdict, while having more negative evaluative expressions than positive ones.

(25) Aside from a couple of **unnecessary** scenes, The Sixth Sense is a low-key **triumph** of mood and menace; the most **shocking** thing about it is how hushed and intimate it is, how softly and quietly it goes about its business of creeping us out. The movie is all **of a piece**, which is probably why the scenes in the trailer, ripped out of context, feel a bit **cheesy**.

If Example 25 is analyzed from the perspective of RST, the EDU 'The Sixth Sense is a low-key triumph of mood and menace' is the nucleus of the highest level RST relation, and thus the central segment of the text. As such, the positive word *triumph* tips the scales in spite of the majority of negative words. A related observation is that evaluative expressions in highly topical sentences get a higher weight.

For movie reviews (as opposed to product reviews), both sentiment analysis and opinion extraction are complicated by the fact that such reviews consist of *descriptive* segments embedded in *evaluative* segments, and *vice versa*. Evaluative expressions in descriptive segments do not contribute as much to the overall sentiment expressed in the review as evaluative expressions in evaluative segments; some of them do not contribute at all (Turney 2002). Consider, for example, *love* in 'I love this movie' and 'The colonel's wife (played by Deborah Kerr) loves the colonel's staff sergeant (played by Burt Lancaster)'; the first but not the second use of the word expresses a sentiment.

From the viewpoint of a flat genre-specific discourse structure, this calls for a distinction of these two kinds of discourse segments, which allows one to assign less weight to evaluative expressions in descriptive segments when calculating the overall opinion in the review (or to ignore them altogether). Pang, Lee and Vaithyanathan (2002) investigated whether such a distinction could be approximated by assuming that specific parts of the review (in particular, its first and last quarter) are evaluative while the rest is devoted to a description of the movie. However, they report that implementing this assumption into their analysis does not improve their results in a significant way.

These observations suggest that discourse analysis (discourse zoning or discourse parsing) has a unique contribution to make to opinion analysis, which is the topic of ongoing work (Voll and Taboada 2007; Taboada, Brooke and Stede 2009).

Voll and Taboada (2007) evaluate the integration of discourse parsing into opinion analysis into their system SO-CAL for automatic sentiment analysis. They compare the results of using only 'discourse-central' evaluative adjectives for assessing the sentiment of movie reviews by SO-CAL against a baseline that uses all these adjectives in the review, and an alternative that only uses evaluative adjectives from topical sentences.

Considering only 'discourse-central' adjectives ignores those adjectives outside the top nuclei of individual sentences, obtained automatically with the discourse parser

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Spade (Soricut and Marcu 2003). This led to a drop in performance, which Voll and Taboada (2007) blame on the discourse parser having only 80% accuracy. An alternative possibility is the way they chose to integrate discourse information into sentiment analysis. It also does not address the task of excluding from sentiment analysis, adjectives from descriptive sections of movie reviews or the problem illustrated in Example 25.

Later work by Taboada *et al.* (2009) uses discourse zoning to distinguish descriptive and evaluative segments of a review. Evaluative expressions from different segments are then weighted differently when the overall opinion of a review is calculated by SO-CAL. This approach is based on experience with the weighing of evaluative expressions within discourse segments, which is used to model the influence of negation, linguistic hedges like *a little bit*, modal expressions like *would*, etc. on the evaluative potential of an expression. They show that the inclusion of information from discourse structure can boost the accuracy of classifying reviews as either positive or negative from 65% to 79%.

In sum, including information on discourse structure into opinion analysis can potentially improve performance by identifying those parts of a discourse whose evaluative expressions are particularly relevant for eventual judgement. Although only a single type of document (short movie reviews) has been studied to date, it is probable that the results of this research will generalize to other kinds of reviews (e.g., for books) as well as to other types of evaluative documents (e.g., client feedback).

This, however, does not exhaust the ways in which discourse structure could contribute to opinion or sentiment analysis. For instance, in comparative reviews (especially of consumer goods), several competing products are evaluated by comparing them feature by feature. Comparisons are often expressed through coherence relations, so recognizing and linking the arguments of these relations could be used to extract all separate judgments about each product. We conclude that research on discourse structure has considerable potential to contribute to opinion analysis, which in our opinion should motivate further attempts to bring together these two threads of research.

4.5 Assessing text quality

Entity chains were introduced earlier (Section 2.2.1) as a feature of Topic Structure, and then as a feature used in algorithms for Topic Segmentation (Section 3.1). Here we briefly describe their use in assessing the naturalness and coherence of automatically generated text.

Barzilay and Lapata (2008) were the first researchers to recognize the potential value of entity chains and their properties for assessing text quality. They showed how one could learn patterns of entity distribution from a corpus and then use the patterns to rank the output of statistical generation. They represent a text in the form of an *entity grid*, a two-dimensional array whose rows correspond to the sequence of sentences in the text, and whose columns correspond to discourse entities evoked by noun phrases. The contents of a grid cell indicate whether the column entity

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

	Pinochet	London	October	Surgery	Arrest	Extradition	Warrant	Judge	Thousands	Spaniards	Hearing	Fate	Balance	Scholars
1	**S**	**X**	**X**	–	–	–	–	–	–	–	–	–	–	–
2	**S**	–	–	**X**	–	–	–	–	–	–	–	–	–	–
3	–	–	–	–	**S**	**X**	**X**	**O**	–	–	–	–	–	–
4	**S**	–	–	–	–	–	–	–	**O**	**O**	–	–	–	–
5	**S**	–	–	–	–	–	–	–	–	–	**O**	**X**	**X**	–
6	–	–	–	–	**O**	–	–	–	–	–	–	–	–	**S**

Fig. 8. Entity grid.

appears in the row sentence and if so, in what grammatical role: as a grammatical subject (S), a grammatical object (O), some other grammatical role (X), or absent (-). A short section of an entity grid is shown in Figure 8.

Inspired by the Centering Theory (Grosz, Joshi and Weinstein 1995), Barzilay and Lapata (2008) consider patterns of *local entity transitions*. A local entity transition is a sequence $\{\mathbf{s}, \mathbf{o}, \mathbf{x}, \mathbf{-}\}^n$ that represents entity occurrences and their syntactic roles in n successive sentences. It can be extracted as a continuous subsequence from a column in the grid. Since each transition has a certain probability in a given grid, each text can be viewed as a distribution over local entity transitions. A set of coherent texts can thus be taken as a source of patterns for assessing the coherence of new texts. Coherence constraints are also modeled in the grid representation implicitly by entity transition sequences, which are encoded using a standard feature vector notation: each grid x_{ij} for document d_i is represented by a feature vector

$$\Phi(x_{ij}) = (p_1(x_{ij}), p_2(x_{ij}), \ldots, p_m(x_{ij}))$$

where m is the number of predefined entity transitions, and $p_t(x_{ij})$ is the probability of transition t in grid x_{ij}.

To evaluate the contribution of three types of linguistic knowledge to model performance (i.e., syntax, coreference resolution, and salience), Barzilay and Lapata (2008) compared their model to models using linguistically impoverished representations. Omitting syntactic information is shown to cause a uniform drop in performance, which confirms its importance for coherence analysis. Accurate identification of coreferring entities is a prerequisite to the derivation of accurate salience models, and salience has been shown to have a clear advantage over other methods. Thus, Barzilay and Lapata provide empirical support for the idea that coherent texts are characterized by transitions with particular properties that do not hold for all discourses. Their work also measures the predictive power of various linguistic features for the task of coherence assessment.

In this work, a sentence is a bag of entities associated with syntactic roles. A mention of an entity, though, may contain more information than just its head and syntactic role. Thus, Elsner and Charniak (2008a), inspired by work on coreference resolution, consider additional discourse-related information in referring expressions – information distinguishing familiar entities from unfamiliar ones and salient

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

entities from nonsalient ones. They offer two models which complement Barzilay and Lapata's (2008) entity grid. Their first model distinguishes *discourse-new* noun phrases whose referents have not been previously mentioned in a given discourse from discourse-old noun phrases. Their second model keeps pronouns close to referents with correct number and gender. Both models improve on the results achieved in Barzilay and Lapata (2008) without using coreference links, which are often erroneous because the disordered input text is so dissimilar to the training data. Instead, they exploit their two models' ability to measure the probability of various aspects of the text.

To sum up this section, different NLP applications make use of automated analysis of discourse structure. For this analysis to be of value for applications, they must have access to robust systems for automated discourse analysis. Right now, the most robust systems are ones for linear discourse segmentation, and so these are most widely used in applications of discourse structure. In contrast, the full range of a hierarchical discourse structure is used only in few applications, in particular, text summarizers. Parts of discourse structure that applications take into account are either sentence-level discourse structures, the top level of the structure, or the discourse relations that link specific segments in a discourse.

5 Supporting algorithms and applications

Technology advances through the public availability of *resources* and through *standardization* that allows them to be used simply 'out of the box'. Here we describe discourse resources available in single languages or genres (Section 5.1) and factored discourse resources that integrate multiple levels of annotation (Section 5.2). For recent efforts at standardization, the reader is referred to Petukhova and Bunt (2009) and Ide, Prasad and Joshi (2011).

5.1 Resources

There is a growing number of textual resources annotated with some form of discourse structure. Some of this annotation is intrinsic, as in the topical sub-heading structure of Wikipedia articles and the conventionalized functional sub-heading structure of *structured abstracts* (Section 3.1). The rest of this section describes resources under development in different languages and genres that have been annotated with discourse relations, intentional structure, or both.

5.1.1 English

English has several resources annotated for some form of discourse structure. The earliest is the RST Discourse TreeBank (Carlson *et al.* 2003), which has been annotated for discourse relations (Section 2.2.4) in a framework adapted from Mann and Thompson (1988) that produces a complete tree-structured RST analysis of each text. The RST corpus comprises 385 articles from PDTB (Marcus, Santorini and Marcinkiewicz 1993), and is available from the Linguistics Data Consortium (http://www.ldc.upenn.edu, CatalogEntry=LDC2002T07).

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Also available from the Linguistics Data Consortium is the Discourse Graph Bank (CatalogEntry=LDC2005T08), based on Wolf's PhD thesis and described in Wolf and Gibson (2005). The corpus comprises 135 articles from the AP Newswire and the PDTB, thus partially overlapping with the set of articles annotated in the RST corpus. The style of annotation in the Discourse Graph Bank differs significantly from other resources as annotators were told to annotate all discourse relations that could be taken to hold between a discourse segment and any segment to its left. Each such relation was also annotated with its sense. Unlike in the RST corpus, annotators did not then attempt to link up the resulting structures. The results is a rather flat discourse structure with frequent crossing arcs and nodes with multiple parents. Egg and Redeker (2010) have pointed out that many of these discourse relations are motivated solely by lexical or referential overlap. Webber (2006) makes a similar point in terms of the number of these long-distance relations that are simply instances of ELABORATION, forming *entity chains* (Section 2.2.1).

The largest resource annotated with some form of discourse structure is the PDTB (Prasad *et al*. 2007, 2008), also available from the Linguistics Data Consortium (CatalogEntry=LDC2008T05). The PDTB 2.0 is lexically grounded annotating discourse relations signalled by explicit discourse connectives drawn from syntactically well-defined classes, as well as relations between adjacent sentences that are not signalled by an explicit connective. Annotators were also allowed to note discourse relations signalled by expressions from outside the annotated classes (ALTLEX) or 'entity relations' between adjacent sentences (ENTREL) where the only relation was between an entity mentioned in one sentence and the sentence following. Approximately 40K discourse relations have been annotated in the PDTB 2.0 (approximately 18.5K explicit discourse relations, 16K implicit discourse relations, 5K ENTRELS, and the remainder, ALTLEX or no relation). Work on *Discourse Chunking* (Section 3.2) has used the PDTB 2.0.

5.1.2 German

The Potsdam Commentary Corpus (PCC) (Stede 2004) consists of 170 commentaries from the German regional daily newspaper *Märkische Allgemeine Zeitung*, which have been annotated with part of speech tags, syntactic analyses, rhetorical structure in the style of RST (Mann and Thompson 1988), discourse relations associated with explicit discourse connectives, links from anaphoric and bridging expressions to the antecedents that license them, and information structure (Topic-Comment, Theme-Rheme, and Cognitive Status). Commentaries were chosen so as to have a more interesting rhetorical structure than either narratives or news reports, and a regional periodical was chosen so as to have a simpler lexicon and syntactic structure than commonly found in national periodicals. The main feature of the PCC is the multiplicity of linguistic analyses and its commitment to their simultaneous accessibility via the same search tool.

Other features of the PCC are its use of a format called URML for underspecifying rhetorical structure. This allows alternative competing analyses to be annotated, rather than just a single one. Explicit connectives are annotated using a bespoke

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

tool that highlights all instances of potential connectives, allowing the annotator to remove those tokens that are not functioning as connectives, and then offering heuristic-based suggestions as to the arguments of each connective, which the annotator can always override.

Information about the PCC can be found at http://www-old.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.en.html. Further discussion of the value of having simultaneous access to multiple levels of linguistic and discourse analysis can be found in Section 5.2.

5.1.3 Danish, English, Italian, and Spanish

The Copenhagen Dependency TreeBank (CDT) (Buch-Kromann, Korzen and Müller 2009; Buch-Kromann and Korzen 2010) comprises four parallel treebanks for Danish, English, Italian, and Spanish that have been richly annotated (morphology, syntax, discourse structure, and coreference) and word aligned with the Danish source text for translational equivalence. The Danish source texts consist of 200–250-word extracts from a range of general texts. As of this writing, there are approximately 440 parallel texts for Danish and English in the CDT that have been syntactically annotated, and seventy-two each for Spanish and Italian. Of these, approximately 340 Danish texts have been annotated for anaphora and discourse structure, and seventy-two each for English, Spanish, and Italian.

Both the syntactic and discourse structure annotations take the form of dependency structures. For discourse, CDT annotation consists in linking up the top dependency node of each sentence with the unique word from some other sentences deemed to govern the relation, and labelling the relation between them. Where a discourse connective signals a relation, it is simply noted as a feature on the relation's sense label. The CDT resembles the RST Discourse TreeBank (Section 5.1.1) both in assuming a nucleus-satellite distinction on all discourse relations (Section 2.3.3) and in taking a tree-structured analysis to fully cover the text (Section 2.3.2). It is unique in having both primary and secondary discourse relations, with the latter encoding anaphoric relations (including coreference) and information related to attribution.

The CDT treebank, annotation manual, and relation hierarchy can be downloaded from http://code.google.com/p/copenhagen-dependency-treebank

5.1.4 Turkish

The METU Turkish Discourse Bank (TDB) (Zeyrek and Webber 2008; Zeyrek *et al.* 2009, 2010) aims to annotate a 500K-word sub-corpus of the 2-million word METU Turkish Corpus (Say *et al.* 2004). The sub-corpus contains a wide range of texts – novels, newspaper columns, memoirs, etc. The initial annotation has focussed on discourse relations that are signalled by explicit discourse connectives, realized either as words or affixes. The style of TDB discourse annotation is similar to that of the PDTB, with several interesting differences. The first is that phrasal expressions, such as *buna rağmen* ('despite this'), have been annotated. These consist of one part that refers to the discourse relation and another part that anaphorically refers to *Arg1*.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

While similar adverbials occur in English, these were not systematically annotated in the PDTB. However, not to do so in the TBD would be to miss an important property of Turkish discourse.

Secondly, the TDB distinguishes among different types of modifiers that can occur with connectives, including adverbials, focus particles, and polarity markers. (While modifiers have been noted in the PDTB, they have not been the subject of further analysis.) A third difference arises from the fact that Turkish is a pro-drop and free word-order language, with subjects and objects dropped if discourse's salient and overt subjects and objects are able to appear in initial, medial, or final position in the sentence. For this reason, the TDB also identifies subjects, objects, and any temporal adverbs that are shared by the arguments of a discourse relation, to enable their meaning to be recovered. This feature is also likely to be useful for learning models able to recover the intended sense of an explicit discourse connective.

More information about the project can be found at the TDB web site (http://www.ii.metu.edu.tr/research_group/metu-turkish-discourse-resource-project).

5.1.5 *Hindi*

The Hindi Discourse Relation Bank (HDRB) (Oza *et al.* 2009) aims to annotate a 200K-word corpus, drawn from a 400K-word corpus of news articles from the Hindi newspaper *Amar Ujala* whose sentences have been annotated with syntactic dependencies. The style of HDRB discourse annotation is similar to that of the PDTB, with interesting differences. First, it considers a wider range of explicit signals of discourse relations: In addition to coordinating conjunctions, subordinating conjunctions, and discourse adverbials, the HDRB annotates sentential relativizers (i.e., relative pronouns that link a relative clause to a matrix clause, rather than to a noun phrase) and particles that indicate the emphatic inclusion of verbs into some relevant situation, much like the use of *also* in English. Secondly, *Arg1* and **Arg2** are specific to particular discourse relation senses, rather than associated generically with clause-level syntax. For example, in a CAUSAL relation, **Arg2** will be assigned to the cause and *Arg1* to the effect. (Syntax was the basis for *Arg1*/**Arg2** distinction in the PDTB because arguments were annotated prior to any sense annotation, not for any deeper reason.) Thirdly, following Sweetser (1990) and Knott (2001), the HDRB makes a more principled three-way distinction between the pragmatic senses of discourse relations:

- An EPISTEMIC sense that relates an epistemic conclusion (*Arg1*) to an observation (**Arg2**), as in 'John loved Mary, because he came back.'
- A SPEECH-ACT sense that relates the performance of a speech act (*Arg1*) to a justification for performing it (**Arg2**), as in 'Do you need anything, because I'm going to the supermarket?'
- A PROPOSITIONAL sense that relates a proposition inferred from one argument with the content of the other argument, as in 'I like Iglu, but we can eat wherever you'd prefer' where inferred from 'I like Iglu' is 'we can/should eat there', which is denied by 'we can eat wherever you'd prefer'.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

A comparison of a small amount of HDRB annotation with PDTB annotation by Oza *et al.* (2009) shows that a significantly larger percentage of discourse relations are conveyed explicitly, but through some means other than a conjunction, discourse adverbial, sentence relativizer, or emphatic particle. Further analysis, along with further analysis of the same phenomenon in the PDTB, should lead to greater understanding of the many ways in which discourse relations are conveyed.

5.1.6 Modern Standard Arabic

The Leeds Arabic Discourse TreeBank (LADTB) (Al-Saif and Markert 2010, 2011) is the first principled discourse annotation effort for Arabic. It is based on the Modern Standard Arabic Penn Treebank v.2 (Maamouri and Bies 2004) whose syntactic annotation has been augmented in the style of the PDTB 2.0 (Section 5.1.1) by annotating the explicit discourse connectives, their arguments, and their senses.

The LADTB comprises 534 news articles (news wire from the Agence France Press) annotated by two judges, with disagreements reconciled by an independent adjudicator. It contains 6,328 annotated tokens of explicit discourse connectives from eighty different types (including clitics, prepositions, and modified forms of basic connectives). Arguments are, for the most part, clauses or anaphoric references to clauses, though as in Turkish, discourse connectives can take a more noun phrase-like unit as an argument (called in Arabic, *Al-Masdar*). The developers note both a greater frequency of explicit connectives than in English text (in particular, between adjacent sentences) and a higher degree of sense ambiguity than with English connectives (Pitler *et al.* 2008). This holds especially true of the rhetorical use of *w* ('and') at the beginning of paragraphs.

As of this writing, the LADTB had not yet been released.

5.1.7 Czech

For the next version of the Prague Dependency TreeBank (PDTB), PDT 3.0, a new layer of annotation is being created (Mladová, Šárka and Hajičová 2008) that will capture the connective relations in discourse, both within a sentence (through coordinating and subordinating relations) and across the sentence boundary. While this layer is inspired in part by the PDTB 2.0, it differs in two main ways: (1) Because all sentences of a text are interlinked into a form of dependency 'megatree' by some type of 'intersentential' relation, annotators can select nodes of this megatree as arguments of discourse connectives, and not just (possibly discontinuous) spans of one or more clauses; (2) because some syntactic constructions linked to connective relations have already been annotated intra-sententially (i.e., through clausal co-ordination, clausal subordination, and 'reference to the preceding context'), these are taken directly over to the discourse annotation layer. As of this writing, the PDT 3.0 has not been released, so only the intra-sentential annotation related to discourse structure is available for current use.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

5.1.8 Dutch

The most recent language-specific effort to annotate discourse structure involves a corpus of eighty Dutch texts being annotated for discourse structure and for relational and lexical cohesion (van der Vliet *et al.* 2011). Because texts from different genres can evince different structures (Section 2.2.2) which is worth characterizing in more detail, the corpus includes forty expository texts (twenty articles on astronomy from an on-line encyclopedia and twenty from a popular science web site) and forty persuasive texts (twenty fund-raising letters and twenty commercial advertisements). Discourse structure is annotated in the style of the RST corpus (Section 5.1.1), with the omission of two problematic relations (ATTRIBUTION and SAME) that were not part of the original proposal for RST (Mann and Thompson 1988). Additional genre-specific structures are being annotated as well. The annotation of relational cohesion involves all lexical and phrasal elements that signal coherence relations at either locally or globally, while the annotation of lexical cohesion involves both repetition (full or partial) and standard semantic relations between nouns, verbs, adjectives, and adverbs.

5.1.9 Discourse annotation of scientific research papers

This and the next section describe discourse-annotated corpora from specific genres: in this section, scientific research papers in English, and in the next, spoken conversations in Italian.

The Biomedical Discourse Relation Bank (BioDRB) (Prasad *et al.* 2011) annotates explicit and implicit discourse relations in the style of the PDTB 2.0 (Section 5.1.1) in a twenty-four-article subset of the GENIA corpus of biomedical research papers (Kim *et al.* 2003) that has also been annotated for coreference and citation relations (Agarwal, Choubey and Yu 2010). Of the 5,859 annotated discourse relations in the BioDRB, 2,636 (45%) involve explicit discourse connectives, 3,001 (51.2%) involve implicit discourse relations, and 193 (3.3%) are annotated as AltLex.

While the BioDRB adheres to most of the PDTB 2.0 annotation conventions, it also allows implicit relations to hold between *non-adjacent arguments* within the same paragraph. (A similar convention will be adopted in the next version of the PDTB.) The BioDRB also has a flatter sense hierarchy, since most of the top-level categories in the PDTB 2.0 were found too broad to be useful. It also includes some senses that were found to be both important and common in the BioDRB, while dropping all 'pragmatic' senses in the PDTB 2.0, except those corresponding to CLAIM and JUSTIFICATION, here subtypes of CAUSE.

Just as the distribution of fine-grained functional categories has been found to differ among different sections of scientific research papers (Section 3.1.2), the distribution of sense relations was found to vary considerably across different functional sections of the BioDRB research papers (Section 3.1). Prasad *et al.* (2011) speculate that functional section may thus be a useful feature in classifying the sense of discourse relations. The BioDRB is available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

The ART Corpus (Liakata *et al.* 2010) is a corpus of 265 papers from physical chemistry and biochemistry, each of whose sentences have been annotated (by one or more experts) with a functional label indicating the component of scientific investigation that it relates to: MOTIVATION, GOAL, OBJECT, METHOD, EXPERIMENT, OBSERVATION, RESULT, CONCLUSION, HYPOTHESIS, MODEL, or BACKGROUND. This *Core Scientific Concept* (or *CoreSC*) annotation scheme also indicates whether the information is about a previous concept or event (OLD) or the concept or event now being reported on (NEW) – e.g., METHOD-NEW_ADVANTAGE: an advantage of the current method, MODEL: a statement about a theoretical model or framework. The corpus can be downloaded from http://www.aber.ac.uk/en/ns/research/cb/projects/art/art-corpus/

The AZ-II Corpus (Teufel, Siddharthan and Batchelor 2009; Liakata *et al.* 2010) is a corpus of sixty-one articles from the Royal Society of Chemistry, each of whose sentences has been annotated (by multiple expert-trained amateurs) with a label indicating its rhetorical or argumentational function and connections it makes between the current work and work in a cited paper. (The annotation scheme is called *AZ-II*, an extension of the approach of *argumentative zoning* earlier described in Teufel and Moens (2002).) In contrast with the ART Corpus, it treats a scientific paper more as a reasoned argument supported by experimental evidence than as a report of a scientific investigation. Thirty-six papers belonging to both corpora have been annotated with both AZ-II and CoreSC annotation. Liakata *et al.* (2010) contains a detailed description of relations revealed between the two annotation schemes and their complementarity.

5.1.10 Discourse annotation of dialogue

Even though we have avoided discussion of spoken dialogue in this survey, a corpus of spoken dialogue annotated for discourse relations in much the same way as the corpora mentioned above may be of interest to readers. Such a corpus of 500 Italian conversations about computer troubleshooting is described by Tonelli *et al.* (2010), recorded at the help-desk facility of the Consortium for Information Systems of Piedmont Region. Within the corpus, all conversations have been segmented at the turn level and annotated with predicate–argument structures, dialogue acts, and concepts and relations drawn from a predefined domain attribute ontology.

In addition, sixty of the conversations have also been annotated with discourse relations in the style of the PDTB 2.0 (Section 5.1.1), both within and across turn boundaries. Of interest is the way this annotation differs from that of the PDTB 2.0. First, because spoken dialogue is more fragmented than written text, annotation of implicit relations could not be limited to adjacent sentences: Here they are annotated wherever they are taken to occur. As for the hierarchy of sense labels, even technical troubleshooting dialogues have more pragmatically motivated relations between utterances than are found in formal written discourse. Thus, the sense hierarchy here is like the Hindi Discourse Relation Bank (Section 5.1.5) in making a principled, three-way distinction among pragmatic senses of connectives. Finally, motivated by the demands of dialogue, one implicit relation was added to the sense hierarchy to annotate instances of REPETITION.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Analysis of the sixty dialogues annotated with discourse relations shows a much greater proportion of explicit connectives than in the PDTB 2.0, even with annotators having been given the freedom to identify implicit discourse relations between nonadjacent units (65.5% explicitly marked discourse relations as compared with 45.8% in the PDTB). Whether this is a feature of Italian or a feature of dialogue has yet to be determined. Other types of analyses that can be carried out on discourse relations in the context of spoken dialogue are described in Petukhova, Préevot and Bunt (2011).

5.2 Factored discourse structure

Many researchers have noted that annotating corpora with information about discourse structure is both demanding and time-consuming. Inter-rater agreement is always an issue, as the task is difficult to fully characterize with a 'long tail' of cases not previously encountered and no independently verifiable truth.

While this problem is partially due to the high level of textual organization of some types of discourse structure, it also follows from assuming a single type of discourse structure fully covering every part of the discourse. For RST, Stede (2008a) presents cases in which this restriction can lead to situations in which the choice between several equally plausible candidate annotations is arbitrary. Some of these situations look like local ambiguities that should be resolvable within larger textual contexts or suggest augmentations of the original RST definitions.

However, other situations indicate genuine problems for monostratal discourse analyses, e.g., the question of how to attach discourse-final evaluative marks. It may not be clear that the marks refer to which part of the preceding discourse (formally, the first argument of the involved EVALUATION relation). Allowing for underspecification in the description might go some way in alleviating the problem but does not offer a principled way of resolving it.

Another problem noted by Stede (2008a) is that RST structures are a conflation of subject-matter (informational) and presentational (intentional) relations. While this looks as if it could motivate conflicting analyses, in practice it does not, because the *most plausible* analysis is chosen in the context of the larger discourse. However, this means that additional relations between discourse segments can get lost. Example 26 from Moore and Pollack (1992) shows a text that instantiates a subject-matter relation (CONDITION – i.e., the conditions under which (c) holds) and a presentational one (MOTIVATION – the desires motivating (a)):

(26) (a) Come home by five o'clock. (b) Then we can go to the hardware store. (c) That way we can finish the bookshelves tonight.

In order to capture the full breadth of discourse-related information, one could alternatively use a more flexible multi-level analysis that allows annotating different aspects of discourse structure simultaneously. Such ideas motivated by Moore and Pollack (1992) were taken up by Moser and Moore (1996), Knott *et al.* (2001), and Poesio *et al.* (2004), and have led to large-scale multi-level analysis and annotation. For instance, the PCC (Stede 2004) includes for example annotations of coreference and discourse connectives and their arguments. Similarly, the PDTB (Prasad *et al.*

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

2007, 2008) was designed to take advantage of syntactic annotation in the Penn TreeBank and any other related annotation of the *Wall Street Journal* corpus – for example, the semantic role labelling in *PropBank* (Kingsbury and Palmer 2002) and the annotation of entity coreference in *OntoNotes* (Hovy *et al.* 2006). In fact, discourse adverbials, such as *after that*, and *in this case* (Section 3.2), were explicitly not annotated in the PDTB to conserve resources under the assumption that a comprehensive annotation of event reference (resolving the demonstrative pronouns *this* and *that* and demonstrative noun phrases) would also cover these types of discourse adverbials.

Recent recoding of part of the PCC abandons the attempt to capture discourse structure in a *single* description, instead dividing it among several layers or *levels* of annotation, starting with coreference (*referential structure*). The second level treats *thematic structure*, encoding the topics and subtopics of the discourse and the way in which they are related to each other. The third level annotates *conjunctive relations*, which are explicitly signalled by discourse connectives. A final level of *intentional structure* aims to encode the speaker's intentions in terms of *speech acts* such as 'stating an option' or 'making a suggestion'. This combines into more comprehensive argument structures for persuasive texts via relations that include 'encourage acting' (≈MOTIVATION in RST) and 'ease understanding' (≈BACKGROUND).

While neither *referential structure* nor *conjunctive relations* and *intentional structure* need cover every discourse segment, discourse segments can still be simultaneously in a conjunctive and an intentional relation, e.g., in Example 26: On the level of conjunctive relations, *then* introduces CONDITION, while the intentional level might be reconstructed as 'encourage acting' (≈MOTIVATION).

To conclude, multi-level annotation allows a much more flexible annotation of discourse structure because it does not force one either to regard discourse structure as a unified phenomenon or to select among conflicting structural properties of a discourse. It is also practical through its potential for reducing the bottleneck of discourse resource creation by distributing the annotation effort and sharing the results. These positive aspects of multi-level annotation are not restricted to discourse structure, and have motivated several efforts to either add new levels to existing annotations (Burchardt *et al.* 2006) or to merge independently developed annotations (Pustejovsky *et al.* 2005; Bex and Verheij 2010). While the latter seems preferrable, it calls for either standardization of form (Ide *et al.* 2011) or content (Petukhova and Bunt 2009) or both, or sophisticated ways of integrating information from the resources involved (Pustejovsky *et al.* 2005; Chiarcos *et al.* 2008). Whatever method is used, aggregating and integrating information from different annotations brings with it the chance of investigating interdependencies between different linguistic levels in novel ways – within discourse structure as well as between other linguistic strata.

6 Future developments

In this section, we comment on two research directions in which we predict discourse structures playing a larger role.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

6.1 Discourse structures and dialogue structure

While this survey has focussed on discourse structures, some newly evolving forms of written discourse, such as blogs, tweets, and on-line forums, involve written contributions from often multiple participants. These forms of written discourse display features of dialogue as well as discourse. Dialogue structure is commonly described in terms of *dialogue acts* (semantic content and the communicative function(s) it is meant to serve) and their *dependency* (or not) on particular previous dialogue acts that their communication function is directed toward (Bunt *et al.* 2010).[14] We foresee researchers in the future exploiting discourse structures (topic structure, high-level functional structure, eventuality structure, and discourse relations), dialogue structure and syntheses thereof in applications involving new forms of interactive written discourse.

This has already begun to happen. The 2009 workshop on *Events in Emerging Text Types* reported on properties of blog entries and blog threads relevant to summarizing such texts. More recently, Wang *et al.* (2011) applied a combination of discourse topic similarity and dialogue analysis to understanding the fine structure of threads in on-line fora. More specifically, within an on-line forum, *posts* (i.e., individual contribution) are connected into *threads* by their 'reply-to' links. A *thread* is thus a unit of discourse similar to research papers, essays, and news reports mentioned elsewhere in this survey. As in both discourse and dialogue, there are richer connections between *posts* than are evident from explicit 'reply-to' links alone, and researchers are beginning to use knowledge of both discourse and dialogue structures to elucidate them (Elsner and Charniak 2008b; Wang *et al.* 2011). Among the goals of this work are to allow richer visualization of thread structure and to identify particularly useful posts (separate from relying on other users to annotate contributions with whether – and if so, how much – they 'like' them).

6.2 Discourse structures and statistical machine translation

Statistical machine translation (SMT) is a second area that we foresee benefitting from the use of discourse structures. Early research on machine translation recognized the importance of one aspect of discourse – correctly translating pronouns so that the same entity was referenced in the target text as in the source. This was attempted through rule-based methods (Mitkov 1999). The only aspect of discourse structure that received any attention was coherence relations (Section 2.2.4), where Marcu, Carlson and Watanabe (2000) suggested that coherence relations might provide an appropriate unit for translation. This is because the intra-sentential and/or inter-sentential realization of coherence relations differs systematically between languages in ways that cannot be predicted from the content of individual utterances alone. However, the only attempt to cash out this insight was a small proof-of-concept system (Ghorbel, Ballim and Coray 2001).

[14] Bunt and colleagues (2010) provides a comprehensive bibliography of work on dialogue structures.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Ten years on, now within the framework of SMT, attention is again turning to discourse phenomena, and exploiting the potential benefits of SMT to recognize discourse structure. Here we briefly note these new efforts, and what more could be done.

6.2.1 Linear discourse segmentation and SMT

In Section 3.1, we showed how linear topic segmentation places segment boundaries between adjacent units when their lexical distributions are sufficiently different. Foster, Isabelle and Kuhn (2010) have now shown that one can exploit this to develop multiple source language models for SMT (each targetted to a different topic) along with multiple translation models. The claim is that together these can provide both more accurate (i.e., topic-relevant) translation for known words and a more natural back-off strategy for unknown words. This is just a first attempt, and more can be done.

When SMT is used as a component in machine-aided human translation, it has been observed that propagating corrections made in post-editing a document can improve how the rest of the document is translated (Hardt and Elming 2010). But Hardt and Elming also find that with highly structured documents, such as patents, corrections made to near-by sentences provide more value than corrections further away. From this, they conclude that, given a source text segmented by either topic structure or high-level functional structure, one could propagate corrections to all and only sentences within the same segment. If segmentation is done using topic models (Section 3.1), then one could also weight corrections in proportion to topic distributions.

6.2.2 Relational structure and SMT

A less comprehensive use of discourse relations than as a unit of translation lies in disambiguating ambiguous discourse connectives so that they can be translated correctly. This is needed because connectives do not have the same ambiguities in all languages. For example, while *since* in English can express either an explanation (like *because*) or a temporal relation (like *after*), there is no similarly ambiguous connective in French: *Puisque* in French expresses only the former sense, while *depuis* expresses only the latter.

Phrase-based SMT lacks sufficient context to disambiguate ambiguous discourse connectives based on context alone. While more context is available in syntax-based SMT, phrase-based SMT allows for more context to be considered through a method called *annotation projection*. In *annotation projection*, instead of source language terms being mapped into the target language, they are first annotated with additional information derived through independent analysis of the source text, with the resulting annotated terms being mapped into the target.

In the case of discourse connectives, carrying out sense disambiguation in the source language (Pitler and Nenkova 2009) and then annotating the disambiguated connectives with their sense class could support more accurate translation, as

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

suggested in a recent preliminary work by Meyer (2011). The technical downside to *annotation projection* is the need for more training data, all of which must be annotated, since annotation results in the probability mass associated with a single term in unannotated source texts being divided up among multiple (annotated) terms in annotated source texts. The deeper downside is the many ways in which discourse relations can be translated, not all of which even use an explicit discourse connective. Evidence for this comes from a study of *explicitation* in manual translation (Koppel and Ordan 2011), which showed that translators often make discourse connectives explicit in their target translation that were implicit in the source. For example, *therefore* and *nevertheless* are about twice as frequent in manually translated text as in source text (0.153% vs. 0.287% for *therefore* and 0.019% vs. 0.045% for *nevertheless*), *thus* is nearly three times as frequent (0.015% vs. 0.041%) and *moreover* is over four times as frequent (0.008% vs. 0.035%).

Explicitation leads to another problem in texts used in training translation models for SMT – source-target misalignments. We thus predict that just as syntactic/dependency structure is starting to be used as the unit of translation so that features of the source conveyed through syntactic and/or dependency structure can be preserved through some means in the target, future research will return to the earlier suggestion (Marcu *et al.* 2000) that discourse relations be considered a appropriate unit of translation.

6.2.3 Anaphora resolution and SMT

We mentioned at the start of this section that early work on Machine Translation recognized the importance of translating pronouns so that their intended coreference was maintained. The problem here is that while anaphoric expressions (pronouns and 0-anaphors) are constrained in all languages by their antecedents, the constraints vary. For example, in English the gender of a pronoun reflects the *referent* of the antecedent, while in French, German, and Czech, pronoun gender reflects its lexical *form*:

(27) a. Here's a book. I wonder if **it** is new. (inanimate, neuter referent)
b. Voici un livre. Je me demande si **il** est nouveau. (masculine form)

In the context of phrase-based SMT, recent works using *annotation projection* from English source texts to French (Nagard and Koehn 2010) and to German (Hardmeier and Federico 2010) have shown some benefits. But a study of its use in projecting to a more distant and morphologically richer language (English to Czech) highlights specific problems in both the method and accurate evaluation (Guillou 2011). Because this problem is so pervasive in translation and one that receives so much attention in monolingual LT, we foresee significant future work here in SMT, possibly informed by one or more types of discourse structures.

7 Conclusion

This review has attempted to show the reader the different ways in which discourse is organized and structured; the ways in which these structures manifest themselves in

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

text; different ways in which these structures can be recovered from 'raw' or parsed text; the ways in which discourse structuring can support various applications in LT; the resources available in different languages to help support new discoveries and applications of discourse structure; and what we can foresee in the future.

We believe that work on discourse structure offers both the promise of 'low-hanging fruit' that will improve performance in LT in the short term and the promise of challenging but solvable problems, both within a single language and cross-linguistically, that can enable LT to perform significantly better and more widely in the future.

References

Agarwal, S., Choubey, L., and Yu, H. 2010. Automatically classifying the role of citations in biomedical articles. In *Proceedings of American Medical Informatics Association* (*AMIA*), *Fall Symposium*, Washington, DC, November 13–17, pp. 11–15.

Agarwal, S., and Yu, H. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics* **25**(23): 3174–80.

Al-Saif, A., and Markert, K. 2010. The Leeds Arabic Discourse Treebank: annotating discourse connectives for Arabic. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May 17–23.

Al-Saif, A., and Markert, K. 2011. Modelling discourse relations for Arabic. In *Proceedings of Empirical Methods in Natural Language Processing*, Edinburgh, Scotland pp. 736–47.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Boston MA: Kluwer.

Asher, N., and Lascarides, A. 2003. *Logics of Conversation*. Cambridge, UK: Cambridge University Press.

Baldridge, J., Asher, N., and Hunter, J. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft* **26**: 213–39.

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, pp. 10–17.

Barzilay, R., and Lapata, M. 2008. Modeling local coherence: an entity-based approach. *Computational Linguistics* **34**(1): 1–34.

Barzilay, R., and Lee, L. 2004. Catching the drift: probabilistic content models with applications to generation and summarization. In *Proceedings of the 2nd Human Language Technology Conference and Annual Meeting of the North American Chapter*, Boston, MA, USA, pp. 113–20. Stroudsburg, PA: Association for Computational Linguistics.

Bestgen, Y. 2006. Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics* **32**(1): 5–12.

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

Buch-Kromann, M., and Korzen, I. 2010 (July). The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden, July 15–16, pp. 127–31.

Buch-Kromann, M., Korzen, I., and Müller, H. H. 2009. Uncovering the 'lost' structure of translations with parallel treebanks. In F. Alves, S. Göpferich, and I. Mees (eds.), *Copenhagen Studies of Language: Methodology, Technology and Innovation in Translation Process Research*, pp. 199–224. Copenhagen Studies of Language, vol. 38. Frederiksberg, Denmark: Copenhagen Business School.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. 2010. Towards an

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

ISO standard for dialogue act annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC 2010*), Valletta, Malta.

Burchardt, A., Frank, A., Erk, K., Kowalski, A., and Padó, S. 2006. SALTO – versatile multi-level annotation tool. In *Proceedings of LREC 2006*, Genoa, Italy.

Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M. 2001. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 98–105. Stroudsburg, PA: Association for Computational Linguistics.

Burstein, J., Marcu, D., and Knight, K. 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**: 32–9.

Callison-Birch, C. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP '08*), Honolulu, HI, USA.

Carlson, L., Marcu, D., and Okurowski, M. E. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*, pp. 85–112. New York: Kluwer.

Chambers, N., and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *Proceedings, Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Columbus, OH, USA, pp. 789–97.

Chen, H., Branavan, S. R. K., Barzilay, R., and Karger, D. 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)*, Boulder, CO, USA, pp. 371–9.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Ldeling, A., Ritz, J., and Stede, M. 2008. A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues* **49**: 271–93.

Choi, F. Y. Y., Wiemer-Hastings, P., and Moore, J. 2001. Latent semantic analysis for text segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP '01*), Pittsburgh, PA USA, pp. 109–17.

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

Clarke, J., and Lapata, M. 2010. Discourse constraints for document compression. *Computational Linguistics* **36**(3): 411–41.

Dale, R. 1992. *Generating Referring Expressions*. Cambridge MA: MIT Press.

Daume III, H., and Marcu, D. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, pp. 449–56.

Do, Q. X., Chan, Y. S., and Roth, D. 2011. Minimally supervised event causality identification. In *Proceedings, Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, pp. 294–303.

Eales, J., Stevens, R., and Robertson, D. 2008. Full-text mining: linking practice, protocols and articles in biological research. In *Proceedings of the BioLink SIG, ISMB 2008*, Toronto, Canada.

Egg, M., and Redeker, G. 2010. How complex is discourse structure? In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC 2010*), Valletta, Malta, pp. 1619–23.

Eisenstein, J., and Barzilay, R. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (EMNLP '08), Honolulu, HI, pp. 334–43.

Elsner, M., and Charniak, E. 2008a. Coreference-inspired coherence modeling. In *Proceedings of ACL-HLT 2008*, Columbus, OH, USA.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Elsner, M., and Charniak, E. 2008b. You talking to me? In *Proceedings of ACL-HLT 2008*, Columbus, OH, pp. 834–42.

Elwell, R., and Baldridge, J. 2008. Discourse connective argument identification with connective specic rankers. In *Proceedings of the IEEE Conference on Semantic Computing* (*ICSC-08*), Santa Clara, CA, USA.

Finlayson, M. 2009. Deriving narrative morphologies via analogical story merging. In *Proceedings, 2nd International Conference on Analogy*, Sofia, Bulgaria, pp. 127–36.

Foster, G., Isabelle, P., and Kuhn, R. 2010. Translating structured documents. In *Proceedings of AMTA*, Atlanta, GA, USA.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*, Sapporo, Japan.

Ghorbel, H., Ballim, A., and Coray, G. 2001. ROSETTA: rhetorical and semantic environment for text alignment. *Proceedings of Corpus Linguistics*, Lancaster, UK, pp. 224–33.

Ghosh, S., Johansson, R., Riccardi, G., and Tonelli, S. 2011b. Shallow discourse parsing with conditional random fields. In *Proceedings, International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, November 8–13.

Ghosh, S., Tonelli, S., Riccardi, G., and Johansson, R. 2011a. End-to-end discourse parser evaluation. In *Proceedings, IEEE Conference on Semantic Computing* (*ICSC-11*), Hong Kong.

Grosz, B., Joshi, A., and Weinstein, S. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics* **21**(2): 203–25.

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

Grosz, B., and Sidner, C. 1990. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack (eds.), *Intentions in Communication*, pp. 417–44. Cambridge MA: MIT Press.

Gu, Z., and Cercone, N. 2006. Segment-based hidden Markov models for information extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July 17–21, pp. 481–8. Stroudsburg PA: Association for Computational Linguistics.

Guillou, L. 2011. *Improving Pronoun Translation for Statistical Machine Translation* (*SMT*). M.Sc. dissertation, University of Edinburgh, Edinburgh, UK.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Sun, L., and Stenius, U. 2010 (July). Identifying the information structure of scientific abstracts. In *Proceedings of the 2010 BioNLP Workshop*, Uppsala, Sweden.

Halliday, M., and Hasan, R. 1976. *Cohesion in English*. Switzerland: Longman.

Hardmeier, C., and Federico, M. 2010. Modelling pronominal anaphora in Statistical Machine Translation. In *Proceedings 7th Int'l Workshop on Spoken Language Translation*, Paris, France, December 2–3, pp. 283–90.

Hardt, D., and Elming, J. 2010. Incremental re-training for post-editing SMT. In *Proceedings of AMTA*, Denver, CO, USA.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *Proceedings, 32nd Annual Meeting of the Association for Computational Linguistics*, Plainsboro, NJ, USA, pp. 9–16.

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

Higgins, D., Burstein, J., Marcu, D., and Gentile, C. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of HLT-NAACL*, Boston, MA, USA, pp. 185–92. Stroudsburg, PA: Association for Computational Linguistics.

Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, Hyderabad, India, pp. 381–8.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Holler, A., and Irmen, L. 2007. Empirically assessing effects of the right frontier constraint. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Conference,* Lagos (Algarve), Portugal, pp. 15–27.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. 2006. OntoNotes: the 90% solution. In *Proceedings, Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 57–60. Stroudsburg, PA: Association for Computational Linguistics.

Ide, N., Prasad, R., and Joshi, A. 2011. Towards interoperability for the Penn Discourse Treebank. In *Proceedings, 6th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (*ISA-6*), Oxford, UK, pp. 49–55.

Kan, M.-Y., Klavans, J., and McKeown, K. 1998. Linear segmentation and segment significance. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. 2003. GENIA corpus – semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(Suppl 1): i180–2.

Kingsbury, P., and Palmer, M. 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evalution* (*LREC*), Spain.

Kintsch, W., and van Dijk, T. 1978. Towards a model of text comprehension and production. *Psychological Review* **85**: 363–94.

Knott, A. 2001. Semantic and pragmatic relations and their intended effects. In T. Sanders, J. Schilperoord, and W. Spooren (eds.), *Text Representation: Linguistic and Psycholinguistic Aspects*, pp. 127–51. Amsterdam: Benjamins.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. 2001. Beyond elaboration: the interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren (eds.), *Text Representation: Linguistic and Psycholinguistic Aspects*, pp. 181–96. Amsterdam: Benjamins.

Koppel, M., and Ordan, N. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting*, pp. 1318–26. Stroudsburg, PA: Association for Computational Linguistics.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., and Webber, B. 2006. Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax? In *Proceedings of the 5th Workshop on Treebanks and Linguistic Theory* (*TLT'06*), Prague, Czech Republic.

Lee, A., Prasad, R., Joshi, A., and Webber, B. 2008. Departures from tree structures in discourse. In *Proceedings of the Workshop on Constraints in Discourse III*, Potsdam, Germany.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Lin, J., Karakos, D., Demner-Fushman, D., and Khudanpur, S. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL Workshop on BioNLP*, Brooklyn, New York, pp. 65–72.

Lin, Z., Ng, H. T., and Kan, M.-Y. 2010 (November). A PDTB-styled end-to-end discourse parser. Technical Report, Department of Computing, National University of Singapore. Available at http://arxiv.org/abs/1011.0835

Lochbaum, K. 1998. A collaborative planning model of intentional structure. *Computational Linguistics* **24**(4), 525–72.

Louis, A., Joshi, A., and Nenkova, A. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pp. 147–56. Stroudsburg, PA: Association for Computational Linguistics.

Louis, A., and Nenkova, A. 2011. General versus specific sentences: automatic identification and application to analysis of news summaries. Technical Report, University of Pennsylvania. Available at http://repository.upenn.edu/cis_reports/

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Maamouri, M., and Bies, A. 2004. Developing an Arabic treebank: methods, guidelines, procedures, and tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, pp. 2–9. Stroudsburg, PA: ACL.

Malioutov, I., and Barzilay, R. 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* (*CoLing-ACL 2006*), Sydney, Australia.

Mandler, J. 1984. *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Hillsdale NJ: Lawrence Erlbaum.

Mani, I. 2001. *Automatic Summarization*. Amsterdam, Netherlands: Benjamins.

Mann, W., and Thompson, S. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**(3), 243–1.

Marcu, D. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of ACL'99*, Maryland, USA, pp. 365–72.

Marcu, D. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics* **26**: 395–448.

Marcu, D., Carlson, L., and Watanabe, M. 2000. The automatic translation of discourse structures. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, Seattle, WA, pp. 9–17.

Marcu, D., and Echihabi, A. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL'02*, Philadelphia, PA, USA.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a large-scale annotated corpus of English: the Penn TreeBank. *Computational Linguistics* **19**: 313–30.

Martin, J. 2000. Beyond exchange: appraisal systems in English. In S. Hunston and G. Thompson (eds.), *Evaluation in Text: Authorial Distance and the Construction of Discourse*, pp. 142–75. Oxford, UK: Oxford University Press.

Maslennikov, M., and Chua, T.-S. 2007. A multi-resolution framework for information extraction from free text. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 592–99. Stroudsburg, PA: Association for Computational Linguistics.

McDonald, R., Crammer, K., and Pereira, F. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL*, Michigan, USA. Stroudsburg, PA: Association for Computational Linguistics.

McKeown, K. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Texts*. Cambridge, UK: Cambridge University Press.

McKnight, L., and Srinivasan, P. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the AMIA Annual Symposium*, Washington DC, pp. 440–44.

Meyer, T. 2011. Disambiguating temporal-contrastive connectives for machine translation. In *Proceedings of the 49th Annual Meeting, Association for Computational Linguistics, Student Session*, pp. 46–51. Stroudsburg, PA: Association for Computational Linguistics.

Mitkov, R. 1999. Introduction: special issue on anaphora resolution in machine translation and multilingual NLP. *Machine Translation* **14**: 159–61.

Mizuta, Y., Korhonen, A., Mullen, T., and Collier, N. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics* **75**: 468–87.

Mladová, L., Šárka, Z., and Hajičová, E. 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*), Marrakech, Morocco.

Moens, M.-F., Uyttendaele, C., and Dumortier, J. 1999. Information extraction from legal texts: the potential of discourse analysis. *International Journal of Human-Computer Studies* **51**: 1155–71.

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

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Moore, J., and Paris, C. 1993. Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics* **19**(4): 651–95.

Moore, J., and Pollack, M. 1992. A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics* **18**(4): 537–44.

Moser, M., and Moore, J. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics* **22**(3): 409–19.

Nagard, R. L., and Koehn, P. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the 5th Joint Workshop on Statistical Machine Translation and Metrics* (*MATR*), Uppsala, Sweden.

Ono, K., Sumita, K., and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings, International Conference on Computational Linguistics* (*COLING*), Kyoto, Japan, pp. 344–48.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., and Joshi, A. 2009. The Hindi Discourse Relation Bank. In *Proceedings of the 3rd ACL Language Annotation Workshop* (*LAW III*), Singapore.

Palau, R. M., and Moens, M.-F. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pp. 98–107. New York: ACM.

Pang, B., and Lee, L. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pp. 115–24. Stroudsburg PA: ACL.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 79–86. Stroudsburg PA: Association for Computational Linguistics.

Paris, C. 1988. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics* **14**(3), 64–78.

Pasch, R., Brausse, U., Breindl, E., and Wassner, U. 2003. *Handbuch der Deutschen Konnektoren*. Berlin, Germany: Walter de Gruyter.

Patwardhan, S., and Riloff, E. 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*), Prague, Czech Republic.

Petukhova, V., and Bunt, H. 2009. Towards a multidimensional semantics of discourse markers in spoken dialogue. In *Proceedings, 8th International Conference on Computational Semantics*, Tilburg, The Netherlands, pp. 157–68.

Petukhova, V., Préevot, L., and Bunt, H. 2011. Multi-level discourse relations in dialogue. In *Proceedings, 6th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (*ISA-6*), Oxford, UK, pp. 18–27.

Pitler, E., and Nenkova, A. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the 47th Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP '09*), Singapore.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. 2008. Easily identifiable discourse relations. In *Proceedings, International Conference on Computational Linguistics* (*COLING*), Manchester, UK.

Poesio, M., Stevenson, R., Eugenio, B. D., and Hitzeman, J. 2004. Centering: a parametric theory and its instantiations. *Computational Linguistics* **30**: 309–63.

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., and Ahn, D. 2004a. A rule-based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, p. 10. Stroudsburg, PA: Association for Computational Linguistics.

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., and Ahn, D. 2004b. Sentential structure and discourse parsing. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Polanyi, L., and Zaenen, A. 2004. Contextual valence shifters. In *Proceedings of AAAI Spring Symposium on Attitude*, Stanford CA, USA, p. 10.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., and Webber, B. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *TAL* (*Traitement Automatique des Langues*) **47**(2): 43–63.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., *et al.* 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Morocco.

Prasad, R., Joshi, A., and Webber, B. 2010a. Exploiting scope for shallow discourse parsing. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC 2010*), Valletta, Malta.

Prasad, R., Joshi, A., and Webber, B. 2010b. Realization of discourse relations by other means: alternative lexicalizations. In *Proceedings, International Conference on Computational Linguistics* (*COLING*). Stroudsburg, PA: Association for Computational Linguistics.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. 2011. The Biomedical Discourse Relation Bank. *BMC Bioinformatics* **12**(188): 18. http://www.biomedcentral.com/1471-2015/12/188

Propp, V. 1968. *The Morphology of the Folktale*, 2nd ed. Austin TX: University of Texas Press. Publication of the American Folklore Society, Inc., Bibliographical & Special Series.

Purver, M. 2011. Topic segmentation. In: G. Tur and R. de Mori (eds.), *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Hoboken NJ: Wiley. Chapter 11, doi:1002/9781119992691.ch11.

Purver, M., Griffiths, T., Körding, K. P., and Tenenbaum, J. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings, International Conference on Computational Linguistics (COLING) and the Annual Meeting of the Association for Computational Linguistics*, pp. 17–24. Stroudsburg, PA: Association for Computational Linguistics.

Pustejovsky, J., Meyers, A., Palmer, M., and Poesio, M. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *CorpusAnno '05: Proceedings of the Workshop on Frontiers in Corpus Annotations II*, pp. 5–12. Stroudsburg, PA: Association for Computational Linguistics.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., *et al.* 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics* **76**(2–3): 195–200.

Rumelhart, D. 1975. Notes on a schema for stories. In D. Bobrow and A. Collins (eds.), *Representation and Understanding: Studies in Cognitive Science*, pp. 211–36. New York: Academic Press.

Sagae, K. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of IWPT 2009*, Paris, France.

Sagae, K., and Lavie, A. 2005. A classifier-based parser with linear run-time complexity. In *Proceedings of IWPT 2005*, Vancouver, British Columbia.

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. 2004. Development of a corpus and a treebank for present day written Turkish. In *Current Research in Turkish Linguistics, 11th International Conference on Turkish Linguistics* (*ICTL 2002*), Eastern Mediterranean University, Northern Cyprus, pp. 183–92.

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

Schilder, F. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering* **8**(3): 235–55.

Sibun, P. 1992. Generating text without trees. *Computational Intelligence*, **8**(1): 102–22.

Soricut, R., and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada.

Sporleder, C., and Lascarides, A. 2008. Using automatically labelled examples to classify rhetorical relations: a critical assessment. *Natural Language Engineering* **14**(3): 369–416.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Stede, M. 2004. The Potsdam Commentary Corpus. In *ACL Workshop on Discourse Annotation*. Stroudsburg, PA: ACL.
Stede, M. 2008a. Disambiguating rhetorical structure. *Research on Language and Computation* **6**: 311–32.
Stede, M. 2008b. RST revisited: disentangling nuclearity. In C. Fabricius-Hansen and W. Ramm (eds.), *Subordination versus Coordination in Sentence and Text*, pp. 33–58. Amsterdam, Netherlands: John Benjamins.
Subba, R., and Eugenio, B. D. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of NAACL '09*, pp. 566–74. Stroudsburg, PA: Association for Computational Linguistics.
Subba, R., Eugenio, B. D., and Kim, S. N. 2006. Discourse parsing: learning FOL rules based on rich verb semantic representations to automatically label rhetorical relations. In *Proceedings of the EACL 2006 Workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy.
Sweetser, E. 1990. *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge, UK: Cambridge University Press.
Taboada, M., Brooke, J., and Stede, M. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of SIGDIAL 2009*, London, UK, pp. 62–70.
Taboada, M., and Mann, W. 2006. Applications of rhetorical structure theory. *Discourse Studies* **8**: 567–88.
Tamames, J., and de Lorenzo, V. 2010. EnvMine: a text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics* **11**: 294.
Teufel, S., and Moens, M. 2002. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics* **28**: 409–45.
Teufel, S., Siddharthan, A., and Batchelor, C. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings, Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 1493–502.
Thione, G., van den Berg, M., Polanyi, L., and Culy, C. 2004. Hybrid text summarization: combining external relevance measures with structural analysis. In *Proceedings of the ACL 2004 Workshop Text Summarization Branches Out*, Barcelona, Spain. Stroudsburg, PA: ACL.
Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC 2010*), Valletta, Malta.
Toolan, M. 2006. Narrative: linguistic and structural theories. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd ed., pp. 459–73. Amsterdam, Netherlands: Elsevier.
Turney, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–24.Stroudsburg, PA: Association for Computational Linguistics.
Uzêda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. 2010. A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing* **6**: 1–20.
van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. 2011. Building a discourse-annotated Dutch text corpus. In S. Dipper and H. Zinsmeister (eds.), *Bochumer Linguistische Arbeitsberichte*, 157–71.
Versley, Y. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Workshop on the Annotation and Exploitation of Parallel Corpora* (*AEPC*), NODALIDA, Tartu, Estonia.
Voll, K., and Taboada, M. 2007. Not all words are created equal: extracting semantic orientation as a function of adjective relevance. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence*, Gold Coast, Australia, pp. 337–46.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press

Walker, M., Stent, A., Mairess, F., and Prasad, R. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research* **30**: 413–56.

Wang, L., Lui, M., Kim, S. N., Nivre, J., and Baldwin, T. 2011. Predicting thread discourse structure over technical web forms. In *Proceedings, Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, pp. 13–25.

Webber, B.. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes* **6**(2): 107–35.

Webber, B. 2006. Accounting for discourse relations: constituency and dependency. In M. Butt, M. Dalrymple, and T. King (eds.), *Intelligent Linguistic Architectures*, pp. 339–60. Stanford, CA: CSLI.

Webber, B. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, Suntec, Singapore.

Wellner, B. 2008. *Sequence Models and Ranking Methods for Discourse Parsing*. PhD thesis, Brandeis University, Waltham, MA, USA.

Wellner, B., and Pustejovsky, J. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing* (*EMNLP-07*), Prague, Czech Republic.

Wolf, F., and Gibson, E. 2005. Representing discourse coherence: a corpus-based study. *Computational Linguistics* **31**: 249–87.

Woods, W. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the AFIPS National Computer Conference*, pp. 457–71. Montvale NJ: AFIPS Press.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Ögel Balaban, H., İhsan Y., and Turan, Ü. D. 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*), Uppsala, Sweden.

Zeyrek, D., Turan, Ü. D., Bozsahin, C., Çakıcı, R., *et al.* 2009. Annotating subordinators in the Turkish discourse bank. In *Proceedings of the 3rd Linguistic Annotation Workshop* (*LAW III*), Singapore.

Zeyrek, D., and Webber, B. 2008. A discourse resource for Turkish: annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources* (*ALR6*), Hyderabad, India.

https://doi.org/10.1017/S1351324911000337 Published online by Cambridge University Press