Discourse Structures and Language Technology

Bonnie Webber

School of Informatics
University of Edinburgh
bonnie@inf.ed.ac.uk

May 12, 2011

Parsing Text?

To understand where Discourse Structure can help LT, we can start by asking: **Where do discourse and LT come in contact?**

Discourse has long been ignored in training and testing **parsers**:

- That the Penn TreeBank (PTB) files are in reverse chronological order is irrelevant for these tasks,
- As would be scrambling the order in which sentences appear in the files!

Although both discourse and sentence structure can vary with genre (eg, *news reports*, *reviews*, *letters*, etc.), parsing a new text benefits more from its **words** having occured in the training corpus than texts from the same genre [Plank & van Noord, 2011].

Summarizing Text?

Features of discourse structure can contribute to selecting "important" sentences in **text summarization** [Marcu 2000; Schilder 2002; Louis, Joshi & Nenkova 2010].

e.g. Sentences whose content plays the discourse role of

- explanation, or
- comment, or
- example

are considered to be **subordinate**, so may be omitted from extractive summaries [Endres-Niggemeyer, 1998].

Summarizing Text?

(1) "Mega or non-mega, we feel the prospectus standards need to be considerably improved," he says. (Implicit = REASON) "Disclosures are very poor in India." [wsj_0629]

(2) "Disclosures are very poor in India." (Implicit = INSTANTIATION) He says the big questions – "Do you really need this much money to put up these investments? Have you told investors what is happening in your sector? . . ." – aren't asked of companies coming to market. [wsj_0629]

⇒ "Mega or non-mega, we feel the prospectus standards need to be considerably improved," he says.

Summarizing Text?

Discourse structure can also be used to help repair the anaphoric and coreferential **chaos** of extractive summarization:

(3) [i] More than 130 bodies are reported to have been recovered after a Gulf Air jet carrying 143 people crashed into the Gulf off Bahrain on Wednesday. [ii] Distraught relatives also gathered at Cairo airport, demanding information. [iii] He also declared three days of national mourning. [iv] He said the jet fell "sharply, like an arrow." [Otterbacher et al, 2002]

$\Rightarrow$ Unlike parsing, discourse structure **can** potentially benefit summarization.

Analysing and Scoring Student Essays?

Discourse structure is a factor in **assessing the quality** of student essays [Burstein et al 2001; 2003]

Good essays that respond to a **prompt** show clear structure:

- Introductory material: segments that provide context for interpreting the thesis, a main idea or the conclusion.
- Thesis: segments that state the writers position and are related to the essay prompt.
- Main idea: segments that assert the authors main message in conjunction with the thesis.
- Supporting idea: segments that support the claims made in the main ideas, thesis statements or conclusions.
- Conclusion: segments that summarize the essays argument.

Analysing and Scoring Student Essays?

If such structure is scrambled or difficult to determine, then the quality of an essay suffers.

Sample Prompt [http://www.ets.org/erater/demo/]

> *Often in life we experience a conflict in choosing between something we want to do and something we feel we should do.*
>
> *In your opinion, are there any circumstances in which it is better for people to do what they want to do rather than what they feel they should do? Support your position with evidence from your own experience or your observations of other people.*

Analysing and Scoring Student Essays?

(4) Throughout our lives, we all find ourselves in a situation at least once, where we have to decide whether we do what we want to or what we feel we should do. This is a very common situation specially among young adults; since we have to decide what we want to make out of our lives. I for instance have to decide to become a lawyer or a doctor. I want to be a lawyer, but I feel I should be a doctor. I can not decide to do what I want or what I think I should since I do not know which is better. [http://www.ets.org/erater/demo/essay_sample_d/]

Thesis (given the prompt)? Main ideas (given the thesis)? Supporting ideas (given the main ideas)? Conclusion??

$\Rightarrow$ Unlike parsing, discourse structure **can** potentially benefit analysing and scoring student essays.

Information Extraction?

Because within a genre, discourse structure predicts where particular information will be found, if present, it can also potentially benefit **information extraction**.

In descriptions of criminal cases, the victim and perpetrator will be found **before** the alleged offenses and court opinion are detailed [Moens et al 1999; 2000].

In a letter, the writer's name comes at the **end** of the text.

Opinion Mining and Sentiment Detection?

Because within a genre, discourse structure predicts how information should be interpreted, if present it can also potentially benefit **opinion mining and sentiment detection**.

e.g, Evaluation expressions

- at the end of a review (ie, summarizing the writer's opinion)
- in a prominent position (eg, paragraph-initial)

are generally better predictors of a writer's overall opinion that those found elsewhere [Taboada et al 2009].

Statistical Machine Translation?

Discourse structure has long been ignored in **SMT**.

But preliminary evidence shows that improvements in translating **anaphoric expressions** (with neither phrase-local or tree-local antecedents) can improve SMT [Le Nagard & Koehn 2010; Hardmeier & Federico 2010].

Because these improvements do not lead to improved Bleu scores other metrics are needed in order to assess them [Hardmeier & Federico 2010].

$\Rightarrow$ Even in **SMT**, discourse structure can deliver potential benefit.

Early Computational Discourse Modelling

If we agree that taking account of discourse structure can help LT, why has it not yet really done so?

Start by looking at the **history** of computational work on discourse modelling.

Early computational work generally assumed that discourse had an underlying **tree structure**, similar to the parse tree of a sentence.

At issue was what the internal nodes of the tree and its other formal properties corresponded to.

Rhetorical Structure Theory (RST)

Rhetorical Structure Theory [Mann & Thomson, 1988] associates a discourse with a tree structure through Context-Free (CF) rewrite rules called **schemas**.

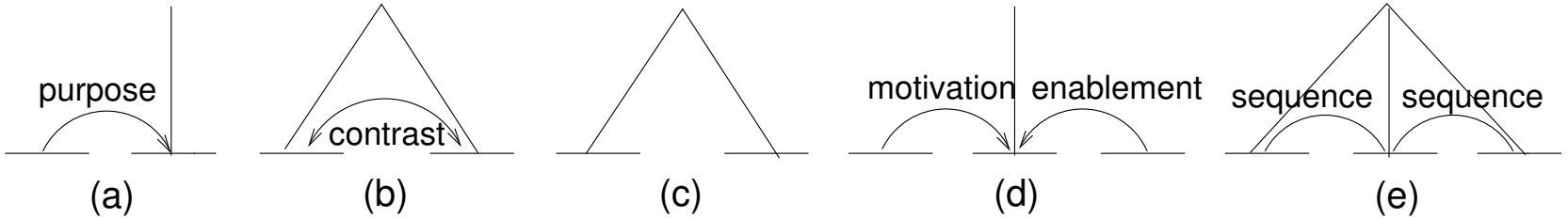
An RST analysis covers the discourse with a tree structure, much as a syntactic parse tree covers a sentence.

Dominance of a node over its children corresponds to a rhetorical relation holding between the text units associated with those child nodes (which project to adjacent text spans).

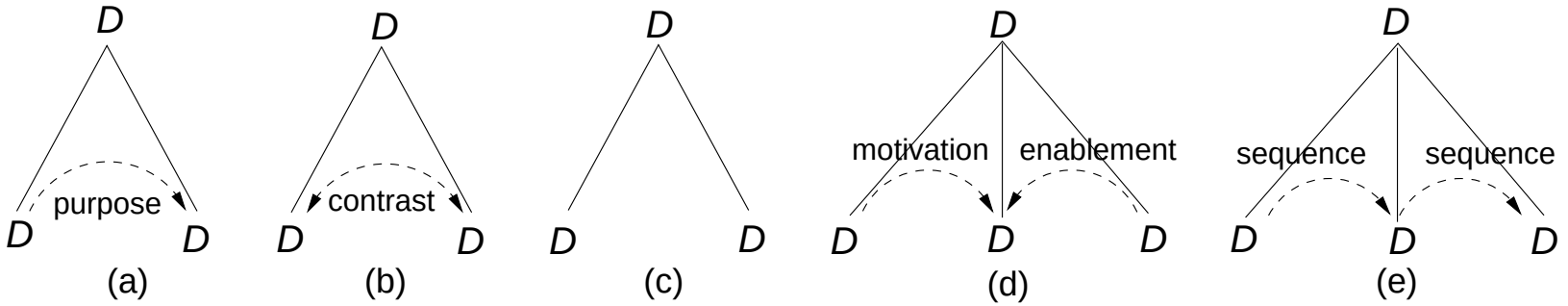
Precedence between nodes corresponds to their order in the text.

RST Schemas as CF rules

RST notation

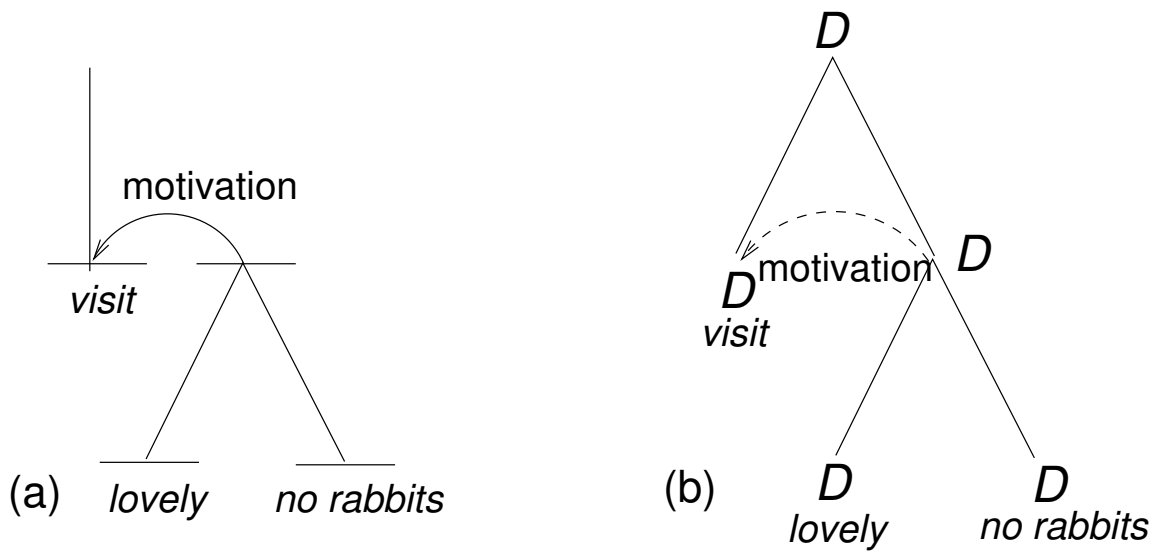


Standard tree notation



Example: RST Analysis

(5) a. You should come visit.
b. Edinburgh is lovely in early fall
c. and there are no rabbits around.



EPICURE [Dale, 1992]

In [Dale, 1992], complex task instructions derive from

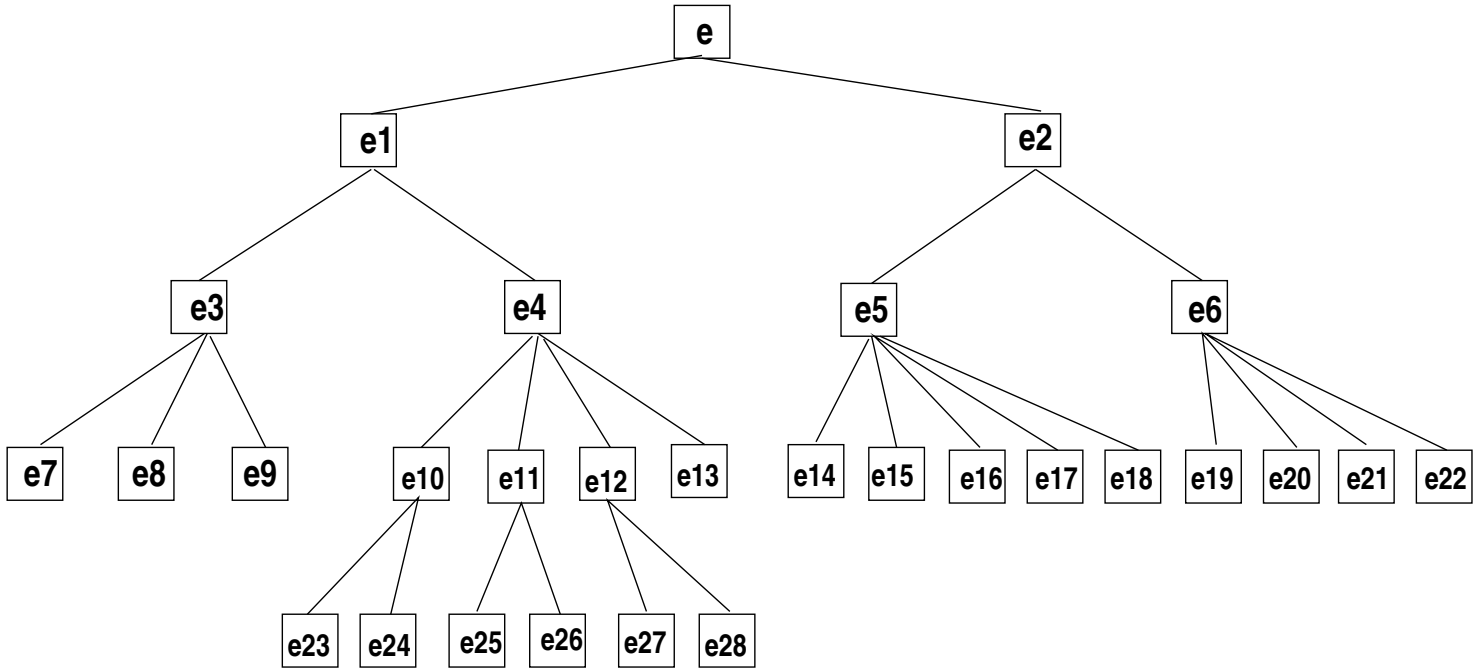
1. a tree structure produced through recursive CF-decomposition of a top-level task into sub-tasks, with **precedence** corresponding to temporal order and **dominance**, to sub-task inclusion;
2. aggregation of similar sisters for realization in a single clause.

Each internal node in the tree corresponded to the next step in a plan to accomplish its parent.

EPICURE [Dale, 1992]

[Make bean soup]	→	[Prep ingredients] [Cook ingredients]
[Prep ingredients]	→	[Prep beans] [Prep veg]
[Prep beans]	→	Soak beans Drain beans Rinse beans
[Prep veg]	→	[Prep onion] [Prep potato] [Prep carrots] Slice celery
[Prep onion]	→	Peel onion Chop onion
[Prep potato]	→	Peel potato Chop potato
[Prep carrots]	→	Scrape carrots Chop carrots
[Cook ingredients]	→	[Stage 1] [Stage 2]
[Stage 1]	→	Melt butter Add veg Saute veg Add beans, stock, milk Simmer
[Stage 2]	→	Liquidise Stir in cream Season Reheat

EPICURE [Dale, 1992]



Generation via top-down LR traversal [Dale, 1992]

Soak, drain and rinse the butter beans. ($e_7 - e_9$)
Peel and chop the onion. ($e_{23} - e_{24}$)
Peel and chop the potato. ($e_{25} - e_{26}$)
Scrape and chop the carrots. ($e_{27} - e_{28}$)
Slice the celery. (e_{13})
Melt the butter. (e_{14})
Add the vegetables. (e_{15})
Saute them. (e_{16})
Add the butter beans, the stock and the milk. (e_{17})
Simmer. (e_{18})
Liquidise the soup. (e_{19})
Stir in the cream. (e_{20})
Add the seasonings. (e_{21})
Reheat. (e_{22})

Intentional Discourse Structure [Grosz and Sidner, 1986]

Grosz & Sidner (1986) posit a tree structure for the *intentional structure* of discourse:

- Nodes correspond to **speaker intentions**.
- **Dominance** in the tree corresponds to the intention of a daughter node supporting that of its parent;
- **Precedence** corresponds to the need to satify an earlier intention before one that follows.

* * * * * * * *

Lost in this early work on the tree structure of discourse was a **linear** model [Sibun, 1992], that seemed to provide a simpler account of certain types of expository text.

Current Computational Discourse Modelling

What has happened since then?

- Continued work on mainly tree-like discourse structures [Asher & Lascarides, 2003; Polanyi et al, 2004] and on some more complex graph structures [Wolf & Gibson, 2005]
- Work on **Topic structure** of discourse
- Work on **Functional structure** of discourse
- Work on **Coherence relations** – "higher-order" pred-arg structure of discourse.

This discourse modelling tends to be data-intensive, reflecting the goal of robust applications.

Topic Structure

Expository text can be viewed as a sequence of topically coherent segments, whose order may become conventionalized over time:

	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	...	...	...

Wikipedia articles about US states

Topic Structure

Being able to recognize topic structure was originally seen as benefitting **information retrieval** [Hearst, 1994; 1997].

Recent interest comes from the potential use of topic structure in **segmenting lectures, meetings or other speech events**, making them more amenable to search [Galley 2003; Malioutov & Barzilay 2006].

Topic Structure

Computational approaches to topic structure and segmentation assume that:

- The topic of each discourse segment relates to the topic of the discourse as a whole (eg, History of Vermont $\rightarrow$ Vermont).
- The only relation holding between sister segments, if any, is pure sequence, although certain sequences may be more common than others (cf. Wikipedia articles).
- The topic of a segment differs from those of its adjacent sisters. (Adjacent spans that share a topic are taken to belong to the same segment.)
- Topic predicts lexical choice, either of all words of a segment or just its content words (ie, excluding "stop-words").

Topic Structure

Making this structure explicit (ie, topic segmentation) uses either

- **semantic-relatedness**, where each segment is taken to consist of words that relate to each other more than to words outside the segment [Hearst 1994, 1997; Choi et al 2001; Bestgen 2006; Galley et al 2003; Malioutov & Barzilay 2006]
- **topic models**, where each segment is taken to be produced by a distinct, compact lexical distribution [Purver et al, 2006; Eisenstein & Barzilay 2008; Chen et al 2009].

[Purver, 2011] contains a useful overview and survey of this work.

Topic Segmentation through Semantic-relatedness

All computational models that use semantic-relatedness for topic segmentation have:

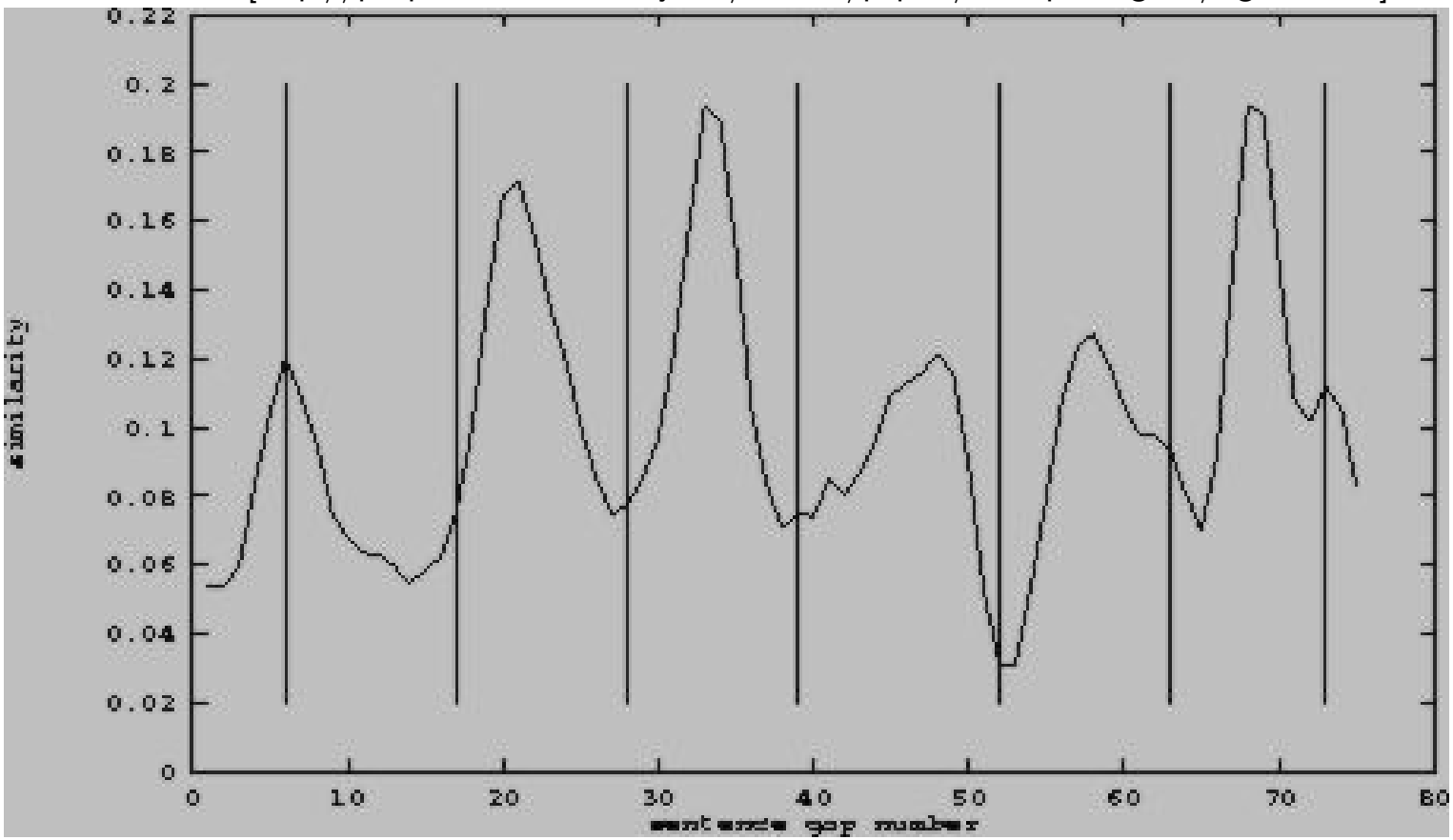
1. a **metric** for assessing the semantic relatedness of terms within a proposed segment;
2. a **locality** that specifies what units within the text are assessed for semantic relatedness;
3. a **threshold** for deciding how low relatedness can drop before it signals a shift to another topic.

TextTiling [Hearst 1994, 1997]

1. **Metric**: *Cosine similarity*, using a vector representation of fixed-length spans (pseudo-sentences) in terms of frequency of word stems (ie, words from which inflection has been removed)
2. **Locality**: Cosine similarity is computed between adjacent spans (and only adjacent spans)
3. **Threshold**: Empirically-determined in order to select where to place segment boundaries.

TextTiling – Computed similarity of adjacent blocks

TextTiling a popular science article. Vertical lines show manually-assigned topic boundaries. Peaks indicate coherency, and valleys, potential breaks between tiles. [http://people.ischool.berkeley.edu/~hearst/papers/subtopics-sigir93/sigir93.html]



Topic Segmentation through Topic Models

Topic segmentation using topic models can take advantage of both

- Features internal to a segment (*Segmental Features*), including words (all words or just content words) and syntax
- Features occurring at segmental boundaries (*Boundary Features*), including discourse cue words (eg, "now", "so", "anyway"), syntax and (in speech) pauses and intonation.

N.B. What cue words are indicating might be better described as **functional structure** than **topic structure**.

Functional structure

Texts within a given genre – eg,

- news reports
- errata
- scientific papers
- letters to the editor of the *New York Times*
- ...

generally share a similar structure, that is independent of topic (eg, sports, politics, disasters; or molecular genetics, radio astronomy, SMT), instead reflecting the **function** played by their parts.

Example: News Reports

Best known is the **inverted pyramid** structure of news reports:

- Headline
- Lead paragraph, conveying **who** is involved, **what** happened, **when** it happened, **where** it happened, **why** it happened, and (optionally) **how** it happened
- Body, providing more detail about who, what, when, ...
- Tail, containing less important information

This is why the first (ie, lead) paragraph is usually the best *extractive summary* of a news report.

Example: Errata

Also recognizable are **errata** – declarations of errors made in previous issue of a periodical and correct versions:

- **Correct statement**
- *Description of error*

(6) **EMPIRE PENCIL, later called Empire-Berol, developed the plastic pencil in 1973.** *Yesterday's Centennial Journal misstated the company's name.* [wsj_1751]

(7) **PRINCE HENRI is the crown prince and hereditary grand duke of Luxembourg.** *An article in the World Business Report of Sept. 22 editions incorrectly referred to his father, Grand Duke Jean, as the crown prince.* [wsj_1871]

Example: Scientific articles/abstracts

Well-known in academia is the multi-part structure of scientific papers (and, more recently, their abstracts):

- **Objective** (aka *Introduction*, *Background*, *Aim*, *Hypothesis*)
- **Methods** (aka *Method*, *Study Design*, *Methodology*, etc.)
- **Results** or *Outcomes*
- **Discussion**
- Optionally, **Conclusions**

N.B. Not every sentence within a section need realise the same function: Fine-grained functional characterizations of scientific papers show them serving a range of functions [Liakata, 2010].

Functional Structure

Automatic annotation of functional structure is seen as benefitting:

- Information extraction: Certain types of information are likely to be found in certain sections [Moens 1999; 2000].
- Extractive summarization: More "important" sentences are more likely to be found in certain sections.
- Sentiment analysis: Words that have an objective sense in one section may have a subjective sense in another [Taboada, 2009].
- Citation analysis: A citation may serve different functions in different sections [Teufel, 2010].

Functional structure

Computational approaches to functional structure and segmentation assume that:

- The function of a segment relates to that of the discourse as a whole.
- While relations may hold between sisters (eg, *Methods* constrain *Results*), only sequence has been used in modelling.
- Function predicts more than lexical choice:
  - indicative phrases such as "results show" ($\rightarrow$ *Results*)
  - indicative stop-words such as "then" ($\rightarrow$ *Method*).
- Functional segments usually appear in a specific order, so either sentence position is a feature used in modelling or sequential models are used.

Functional structure

The internal structure of segments has usually been ignored in high-level functional segmentation [Chung, 2009; Lin et al, 2006; McKnight, 2003; Ruch, 2007].

But given the results of work in fine-grained modelling of functional structure, it is not surprising that Hirohata et al (2008) found significant boundary features:

- Properties of the first sentence of a segment differ from those of the rest (as in 'BIO' approaches to Named Entity Recognition).
- Modelling this leads to improved performance in high-level functional segmentation (ie, 94.3% per sentence accuracy vs. 93.3%).

Labelled biomedical abstracts

Much function-based modelling has been on biomedical text [Chung, 2009; Guo et al, 2010; Hirohata et al, 2008; Liakata et al, 2010; Lin et al, 2006; Mcknight & Srinivasan, 2003; Ruch et al, 2007], where texts with explicitly labelled sections serve as **free training data** for segmenting unlabelled texts.

(8) **BACKGROUND**: Mutation impact extraction is a hitherto unaccomplished task in state of the art mutation extraction systems. . . . **RESULTS**: We present the first rule-based approach for the extraction of mutation impacts on protein properties, categorizing their directionality as positive, negative or neutral. . . . **CONCLUSION**: . . . Our approaches show state of the art levels of precision and recall for Mutation Grounding and respectable level of precision but lower recall for the task of Mutant-Impact relation extraction. . . . [**PMID 21143808**]

Unlabelled biomedical abstracts

(9) We propose two methods for finding similarities in protein structure databases. Our techniques extract feature vectors on triplets of SSEs (Secondary Structure Elements) of proteins. These feature vectors are then indexed using a multidimensional index structure. Our first technique considers the problem of finding proteins similar to a given query protein in a protein dataset. This technique quickly finds promising proteins using the index structure. These proteins are then aligned to the query protein using a popular pairwise alignment tool such as VAST. We also develop a novel statistical model to estimate the goodness of a match using the SSEs. Our second technique considers the problem of joining two protein datasets to find an all-to-all similarity. Experimental results show that our techniques improve the pruning time of VAST 3 to 3.5 times while keeping the sensitivity similar. [PMID 16452789]

Other work on functional segmentation of legal texts [Moens 1999, 2000] and student essays [Burstein et al 2001, 2003].

"Higher-order" pred-arg structures

Discourse also has structure arising from semantic and pragmatic relations that hold between the referents of its clauses.

These resemble relations between the referents of NPs that serve as **args** to a **predicate** conveyed by a verb (**PropBank**) or noun (**NomBank**).

These "higher-order" pred-arg structures (**discourse relations** or **coherence relations**) are usually signalled by a discourse connective

- a conjunction like *because* or *but*,
- a discourse adverbial like *nevertheless* or *instead*.

Discourse Connectives ≠ Discourse Cues

Words like English *so* are ambiguous, with some tokens that serve as discourse connectives and some, as discourse cues.

(10) But C.J.B. Marshall, vicar of a nearby church, feels the fault is in the stairs from the bell tower that are located next to the altar. "**So** crunch, crunch, crunch, bang, bang, bang – here come the ringers from above, making a very obvious exit while the congregation is at prayer," he says. [wsj_0089]

(11) Indeed, Judge O'Brien ruled that "it would be easy to conclude that the USIA's position is 'inappropriate or even stupid,"' but it's the law. So the next step, I suspect, is to try to get the law changed. [wsj_0108]

N.B. Discourse relations may also be signalled in other ways, like English *that means*, *what if*, etc. [Prasad et al, 2008]).

"Higher-order" pred-arg structures

But just as relations between referents can be signalled purely by adjacency — cf. **noun-noun modifier** constructions in English

- human-computer communication
- container ship crane operator courses

coherence relations can be conveyed through adjacency between clauses or sentences (aka *implicit connectives*).

(12) Viewers may not be cheering, either.
Soaring rights fees will lead to an even greater clutter of commercials. [wsj_1057]

"Higher-order" pred-arg structures

But just as relations between referents can be signalled purely by adjacency — cf. **noun-noun modifier** constructions in English

- human-computer communication
- container ship crane operator courses

coherence relations can be conveyed through adjacency between clauses or sentences (aka *implicit connectives*).

(13) Viewers may not be cheering, either. (implicit=REASON)
Soaring rights fees will lead to an even greater clutter of commercials. [wsj_1057]

Resources annotated with Coherence Relations

The Penn Discourse TreeBank is currently the largest resource manually annotated for discourse connectives, their arguments, and the senses they convey [Prasad, 2008].

PDTB Relations	No. of tokens
Explicit	18459
Implicit	16224
...	...

Resources annotated with Coherence Relations

Resources annotated like the PDTB are being created for:

- Modern Standard Arabic [Al-Saif and Markert, 2010]
- Chinese [Xue, 2005, 2010]
- Czech [Mladová et al, 2008]
- Danish and Italian [Buch-Kromann and Korzen, 2010]
- Dutch [van der Vliet et al, 2011]
- German [Stede, 2004, 2008]
- Hindi [Oza et al, 2009]
- Turkish [Zeyrek et al, 2010]

In all these language, coherence relations have two (and only two) arguments.

"Higher-order" pred-arg structures

The coherence relations in a text do not necessarily form a tree:

- A single span may serve as an argument to multiple relations (ie, have incoming edges from different nodes).
- The structure may only be partically connected.

The structure they form may also only be a partial cover of the text.

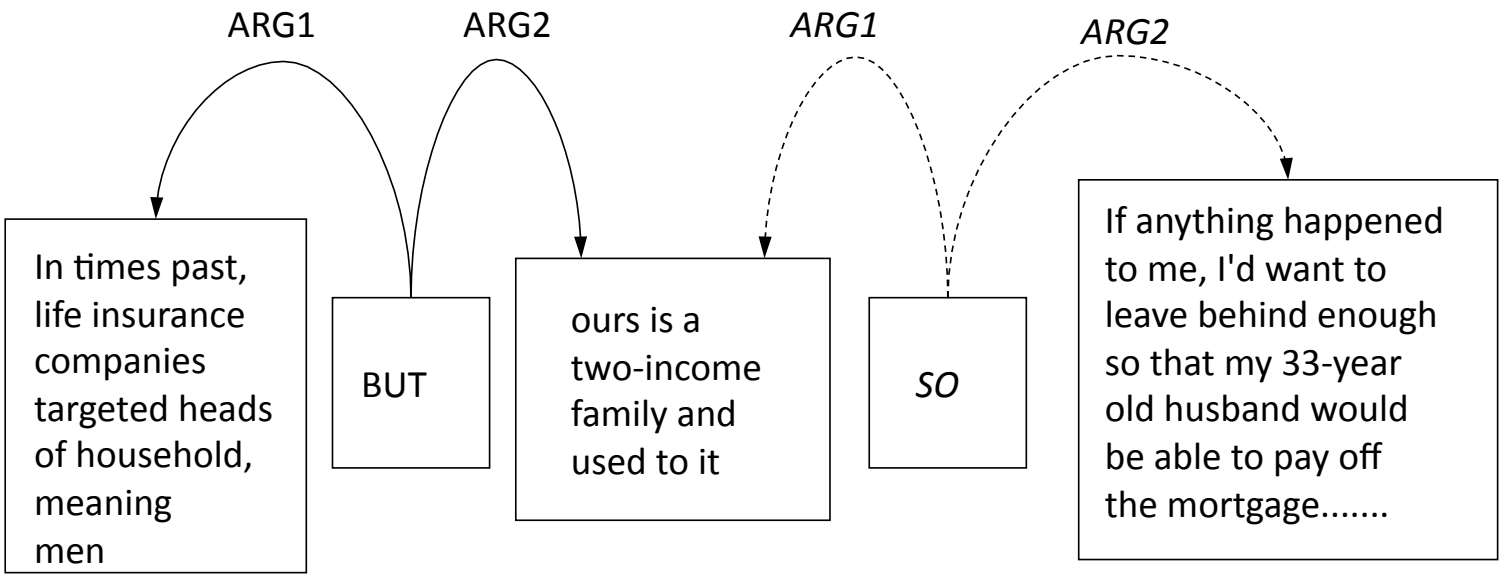
Serving as an arg to multiple relations

(14) In times past, life-insurance salesmen targeted heads of household, meaning men, but ours is a two-income family and accustomed to it. So if anything happened to me, I'd want to leave behind enough so that my 33-year-old husband would be able to pay off the mortgage . . . [Lee et al., 2006]

Serving as an arg to multiple relations

(14) In times past, life-insurance salesmen targeted heads of household, meaning men, but ours is a two-income family and accustomed to it. So if anything happened to me, I'd want to leave behind enough so that my 33-year-old husband would be able to pay off the mortgage ...
[Lee et al., 2006]

Serving as an arg to multiple relations



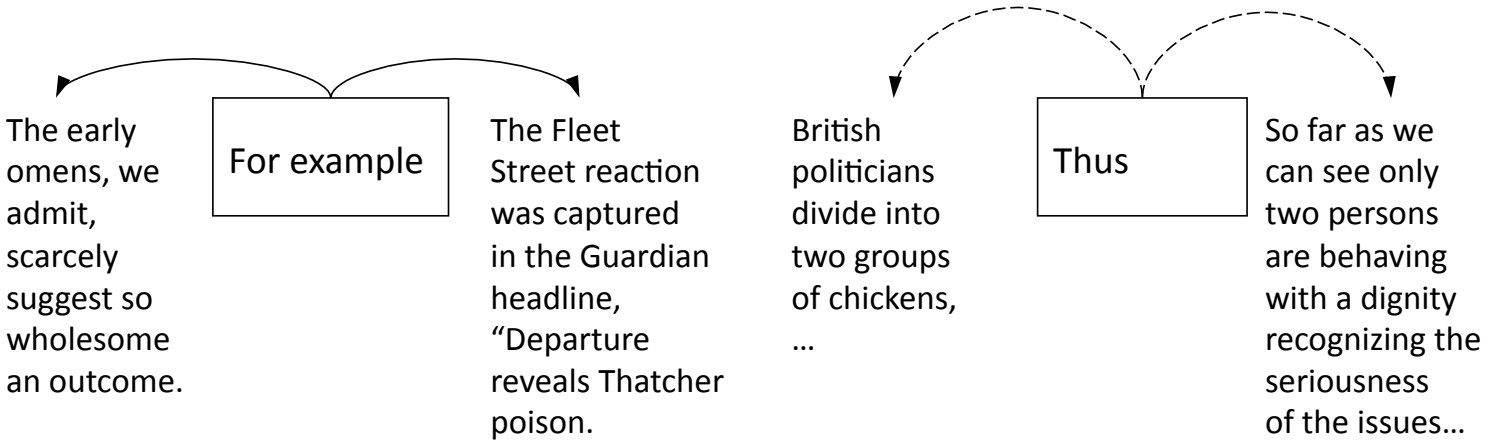
Partial connectivity – Disconnected structures

(15) The early omens, we admit, scarcely suggest so wholesome an outcome. The Fleet Street reaction was captured in the Guardian headline, "Departure Reveals Thatcher Poison." British politicians divide into two groups of chickens, those with their necks cut and those screaming the sky is falling. So far as we can see only two persons are behaving with a dignity recognizing the seriousness of the issues: Mr. Lawson and Sir Alan Walters .... [wsj_0553]

Partial connectivity

(16) The early omens, we admit, scarcely suggest so wholesome an outcome. Implicit=FOR EXAMPLE The Fleet Street reaction was captured in the Guardian headline, "Departure Reveals Thatcher Poison." NoRel British politicians divide into two groups of chickens, those with their necks cut and those screaming the sky is falling. Implicit=THUS So far as we can see only two persons are behaving with a dignity recognizing the seriousness of the issues: Mr. Lawson and Sir Alan Walters .... [wsj_0553]

Partial connectivity



Automatically recognizing coherence relations

Task involves:

- Identifying the evidence for the discourse relation – ie, evidence for the "discourse predicate";
- Identifying the arguments related by that predicate;
- Identifying the sense of the relation.

[Pitler & Nenkova, 2009; Prasad et al. 2008; Wellner & Pustejovsky, 2007; Elwell & Baldridge, 2008; Prasad, Joshi & Webber, 2010]

Future of Discourse Structure and LT

The future promises:

- greater **understanding** of discourse structures
- more accurate, automatic **recognition** of discourse structures
- further **applications** of discourse structures in LT.

Greater understanding of discourse structures

Discourse displays different kinds of structure.

Grosz & Sidner [1986] is an early attempt to jointly model:

- **Linguistic Structure** — a segmental structure on the sequences of utterances
- **Intentional Structure** — a hierarchical structure on the discourse-relevant purpose of each segment and how they relate to one another
- **Attentional Structure** — a stack model of the participants' attentional state as the discourse unfolds.

Although recognizing intentional structure demands significant world and social knowledge, this is clearly easier where intentional structure has become conventionalised to genre-specific **functional structure**.

Joint modelling of discourse structure

Joint modelling is the only way to make sense of some rather strange patterns in the annotation of coherence relations.

Recall **Errata**: One segment describes the error, while the other states what is correct.

(17) [wsj_1751] EMPIRE PENCIL, later called Empire-Berol, developed the plastic pencil in 1973. Yesterday's Centennial Journal misstated the company's name.

Of 23 2-sentence errata in PDTB:

- 11 annotated as forms of Comparison, including Contrast, Contrast.juxtaposition, Contrast.opposition, Concession.expectation
- 10 annotated as EntRel
- 2 annotated as Temporal.asynchronous.succession

Topic Segmentation & Coherence Relations

There is also value in integrating **Topic segmentation** and **Coherence relations**, even though they are independent.

There is a precedent for this:

- Marcu [2000] used topic segmentation to label RST structure beyond that of Elementary Discourse Units (EDUs).
- Schilder [2002] used topic segmentation, instead of RST structure beyond that of EDUs.

Topic Segmentation & HO Pred-Arg Structure

In the other direction, Coherence relations can adjust decisions made in topic segmentation:

(18) Mr. Oxnard observed that the situation in Brazil is also very complicated. On the one hand, Brazil started an ethanol program about 15 years ago to fuel a huge portion of its national fleet of cars and is now committed to this program.

"It has to weigh, on the other hand, the relatively high price of sugar it can earn on the export market in making decisions as to whether to produce sugar or alcohol," Mr. Oxnard said. [wsj_0155]

Topic Segmentation & HO Pred-Arg Structure

No lexical overlap between paragraphs. But para-init "OTOH" relates S2 and S3.

(19) Mr. Oxnard observed that the situation in Brazil is also very complicated. On the one hand, Brazil started an ethanol program about 15 years ago to fuel a huge portion of its national fleet of cars and is now committed to this program.

"It has to weigh, on the other hand, the relatively high price of sugar it can earn on the export market in making decisions as to whether to produce sugar or alcohol," Mr. Oxnard said. [wsj_0155]

N.B. In the WSJ corpus, most occurances of "OTOH" are para-init, thus potentially the start of a new topic.)

Joint Modelling: Functional Segmentation & HO Predicates

ART/CoreSC Corpus contains fine-grained (sentence-level) **functional** annotation of core components of scientific investigations.

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

Annotation categories include:

Background (BAC)	Object (OBJ)	Observation (OBS)
Hypothesis (HYP)	Method (MET)	Result (RES)
Motivation (MOT)	Model (MOD)	Conclusion
Goal (GOA)	Experiment (EXP)	

Joint Modelling: Functional Segmentation & HO Predicates

(20) <s sid="118"><ART atype="GSC" type="Result" ...>
Finally, the overall effect of static correlation is to decrease the atomic size!
</ART></s>

<s sid="119"><ART atype="GSC" type="Result" ...>
The effect is particularly pronounced for transition metal atoms.
</ART></s>

<s sid="120"><ART atype="GSC" type="Observation" ...>
However both atoms with large intra-valence correlation effects (eg V) and those with none (eg Cr) decrease in size on including dynamic correlation effects through introducing PT2 to CASSCF.
</ART></s>

ann1: b402989p_mode2.xml

Conclusion and prediction

Current computational models of discourse structure are tied, more or less, to empirical data:

- mainly tree-like discourse structures and some more complex graph structures
- abstract topics
- conventionalized functions
- "higher-order" pred-arg structure (aka *coherence relations*)
- entity mentions [Knott et al, 2001; Barzilay & Lapata, 2008]

Since natural coherent discourse involves them all, joint modelling may lead to

- better models
- better understanding of discourse

As for modelling blog posts, tweets, etc. . . . ???

References

◦ Nicholas Asher and Alex Lascarides (2003). *Logics of Conversation*, Cambridge University Press.

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

◦ Matthias Buch-Kromann and Iørn Korzen (2010). The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. *Proc. 4th Linguistic Annotation Workshop*, pages 127–131, Uppsala.

◦ Jill Burstein, Daniel Marcu, Slava Andreyev, and Martin Chodorow (2001). Towards automatic classification of discourse elements in essays. *Proc. 39th Annual Meeting of the ACL*, pages 98–105, Toulouse.

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

◦ Harr Chen, S. R. K. Branavan, Regina Barzilay, and David Karger (2009). Global models of document structure using latent permutations. *Proc. NAACL'09*, 371–379.

◦ Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore (2001). Latent Semantic Analysis for text segmentation. *Proc. EMNLP'01*, 109–117.

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

◦ Robert Dale (1992). *Generating Referring Expressions*. Cambridge: MIT Press.

◦ Jacob Eisenstein and Regina Barzilay (2008). Bayesian unsupervised topic segmentation. *Proc. EMNLP*, pages 334–343.

◦ Robert Elwell and Jason Baldridge (2008). Discourse connective argument identication with connective specic rankers. *Proc. IEEE Conference on Semantic Computing (ICSC-08)*, Santa Clara CA.

◦ Brigitte Endres-Niggemeyer (1998). *Summarizing Information*. Springer-Verlag.

◦ Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing (2003). Discourse segmentation of multi-party conversation. *Proc. ACL'03*.

◦ Barbara Grosz and Candace Sidner (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

◦ Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius (2010). Identifying the information structure of scientific abstracts. *Proc. 2010 BioNLP Workshop*, Uppsala, Sweden.

◦ Christian Hardmeier and Marcello Federico (2010). Modelling Pronominal Anaphora in Statistical Machine Translation. *Proc. 7th Int'l Workshop on Spoken Language Translation*.

◦ Marti Hearst (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

◦ Kenji Hirohata, Naoki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka (2008). Identifying sections in scientic abstracts using conditional random fields. *Proc. 3rd Int'l Joint Conf. on Natural Language Processing*, 381–388.

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

◦ Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber (2006). Complexity of dependencies in discourse. *Proc. 5th Int'l Workshop on Treebanks and Linguistic Theories*, Prague.

◦ Maria Liakata, Simone Teufel, Advaith Siddharthan and Colin Batchelor (2010). Corpora for the conceptualisation and zoning of scientific papers. *Proc.7th Conf/ on Language Resources and Evaluation*, Valletta, Malta.

◦ Ronan Le Nagard and Philipp Koehn (2010). Aiding Pronoun Translation with Co-Reference Resolution. *Proc. 5th Joint Workshop on Statistical Machine Translation and Metrics (MATR)*. Uppsala.

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

◦ Annie Louis, Aravind Joshi and Ani Nenkova (2010). Discourse indicators for content selection in summarization. *Proc. SIGDIAL*, Tokyo Japan, 147–156.

◦ Igor Malioutov and Regina Barzilay (2006). Minimum Cut Model for Spoken Lecture Segmentation. *Proc. of ACL/COLING*, Sydney.

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

◦ Daniel Marcu (2000). The rhetorical parsing of unrestricted texts. *Computational Linguistics*, 26(3):395–448.

◦ Larry McKnight and Padmini Srinivasan (2003). Categorization of sentence types in medical abstracts. *Proc. AMIA Annual Symposium*, 440–444.

◦ Lucie Mladová, Šárka Zikánová, and Eva Hajičová (2008). From sentence to discourse: Building an annotation scheme for discourse based on the Prague Dependency Treebank. *Proc. 6th Int'l Conf. on Language Resources and Evaluation*.

◦ Marie-Francine Moens, Caroline Uyttendaele and Jos Dumortier (1999). Information extraction from legal texts: the potential of discourse analysis. *Int'l. Journal of Human-Computer Studies*, 51:1155–1171.

◦ Marie-Francine Moens, Caroline Uyttendaele and Jos Dumortier (2000). Intelligent information extraction from legal texts. *Information & Communications Technology Law*, 9:17–26.

◦ Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma and Aravind Joshi (2009). The Hindi Discourse Relation Bank. *Proc. 3rd Linguistic Annotation Workshop (LAW III)*. Singapore.

◦ Barbara Plank and Gertjan van Noord (2011). Effective Measures of Domain Similarity for Parsing. *Proc. 49th Annual Meeting of the ACL*, Portland OR.

◦ Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn (2004). Sentential structure and discourse parsing. *Proc. ACL 2004 Workshop on Discourse Annotation*.

◦ Matthew Purver (2011). Topic Segmentation. In Gokhan Tur and Renato de Mori (eds.), *Spoken Language Understanding*, Wiley, 2011.

◦ Matthew Purver, Konrad Körding, Thomas Griffiths and Joshua Tenenbaum (2006). Unsupervised Topic Modelling for Multi-Party Spoken Discourse. *Proc. 21st COLING and 44th Annual Meeting of the ACL*, Sydney, pp. 17–24.

◦ Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008). The Penn Discourse TreeBank 2.0. *Proc. 6th Int'l Conference on Language Resources and Evaluation*.

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

◦ Frank Schilder (2002). Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(3):235–255.

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

◦ Manfred Stede (2004). The Potsdam Commentary Corpus. *ACL Workshop on Discourse Annotation*, Barcelona, Spain.

◦ Manfred Stede (2008). Disambiguating rhetorical structure. *Research on Language and Computation*, 6:311–332.

◦ Maite Taboada, Julian Brooke and Manfred Stede (2009). Genre-based paragraph classification for sentiment analysis. *Proc. of SIGDIAL 2009*, pages 62–70, Queen Mary University of London.

◦ Simone Teufel (2010). *The Structure of Scientific Articles*. CSLI Publications, Stanford CA.

◦ Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg and Gisela Redeker (2011). Building a discourse-annotated Dutch text corpus. *Beyond Semantics*, Gottingen, Germany.

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

◦ Nianwen Xue (2005). Annotating discourse connectives in the Chinese Treebank. *Proc. ACL Workshop in Frontiers in Annotation II: Pie in the Sky*. Ann Arbor, Michigan.

◦ Nianwen Xue and Yuping Zhou (2010). Applying Syntactic, Semantic and Discourse Constraints in Chinese Temporal Annotation. *Proc. COLING'2010*, Beijing.

◦ Deniz Zeyrek, Umit Deniz Turan, Cem Bozsahin, Ruket Cakici, Isin Demirsahin, et al (2009). Annotating Subordinators in the Turkish Discourse Bank. *Proc. 3rd Linguistic Annotation Workshop (LAW III)*. Singapore.