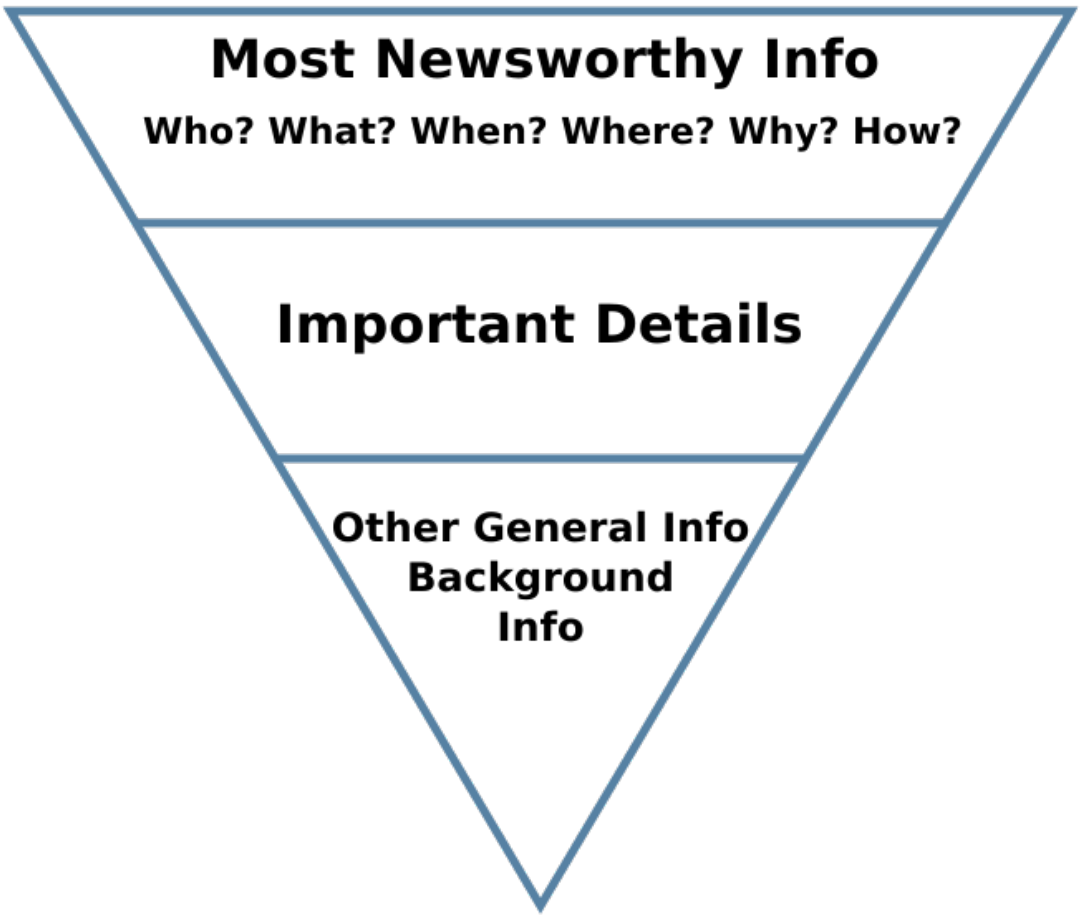
Discourse Structure and Coherence

Discourse segmentation and discourse coherence

1. Discourse segmentation: chunking texts into coherent units. (Also: chunking separate documents)
2. (Local) discourse coherence: characterizing the meaning relationships between clauses in text.

Discourse segmentation examples



(The inverted pyramid design)

Discourse segmentation examples

Clinical Comparison of Full-Field Digital Mammography and Screen-Film Mammography for Detection of Breast Cancer

John M. Lewin[1], Carl J. D'Orsi[2], R. Edward Hendrick[1,3], Lawrence J. Moss[2], Pamela K. Isaacs[1], Andrew Karellas[2] and Gary R. Cutter[4]

[1] University of Colorado Health Sciences Center, 4200 E. 9th Ave., Mail Stop F724, Denver, CO 80262.
[2] University of Massachusetts Medical Center, 55 Lake Ave. N., Worcester, MA 01655.
[3] Northwestern University Medical School, 357 E. Chicago Ave., Chicago, IL 60611.
[4] AMC Cancer Research Center, 1600 Pierce St., Lakewood, CO 80232.

OBJECTIVE. The purpose of this work is to compare full-field digital mammography and screen-film mammography for the detection of breast cancer in a screening population.

SUBJECTS AND METHODS. Full-field digital mammography was performed in addition to screen-film mammography in 6736 examinations of women 40 years old and older presenting for screening mammography at either of two institutions. Two views of each breast were acquired with each technique. The digital and screen-film mammograms were each interpreted independently. In addition to a clinical assessment, each finding was assigned a probability of malignancy for use in receiver operating characteristic analysis. In cases in which the digital and screen-film interpretations differed, a side-by-side analysis was performed to determine the reasons for the discrepancy. With few exceptions, findings detected on either technique were evaluated with additional imaging and, if warranted, biopsy.

RESULTS. Additional evaluation was recommended on at least one technique in 1467 cases. These additional evaluations led to 181 biopsies and the detection of 42 cancers. Nine cancers were detected only on digital mammography, 15 were detected only on screen-film mammography, and 18 were detected on both. The difference in cancer detection is not statistically significant ($p > 0.1$). Digital mammography resulted in fewer recalls than did screenfilm mammography (799 vs 1007, $p < 0.001$). The difference between the receiver operating characteristic curve area for digital (0.74) and screen-film (0.80) mammography was not significant ($p > 0.1$). Reasons for discrepant interpretations of cancer were approximately equally distributed among those relating to lesion conspicuity, lesion appearance, and interpretation.

CONCLUSION. No significant difference in cancer detection was observed between digital mammography and screen-film mammography. Digital mammography resulted in fewer recalls than did screen-film mammography.

(Pubmed highly structured abstract)

Discourse segmentation examples

Identification of Genes Required for the Function of Non-Race-Specific mlo Resistance to Powdery Mildew in Barley

A. Freialdenhoven, C. Peterhansel, J. Kurth, F. Kreuzaler and P. Schulze-Lefert
Rheinisch-Westfalische Technische Hochschule Aachen, Department of Biology I, Worringer Weg 1, D-52074 Aachen, Germany

Recessive alleles (mlo) of the Mlo locus in barley mediate a broad, non-race-specific resistance reaction to the powdery mildew fungus Erysiphe graminis f sp hordei. A mutational approach was used to identify genes that are required for the function of mlo. Six susceptible M2 individuals were isolated after inoculation with the fungal isolate K1 from chemically mutagenized seed carrying the mlo-5 allele. Susceptibility in each of these individuals is due to monogenic, recessively inherited mutations in loci unlinked to mlo. The mutants identify two unlinked complementation groups, designated Ror1 and Ror2 (required for mlo-specified resistance). Both Ror genes are required for the function of different tested mlo alleles and for mlo function after challenge with different isolates of E. g. f sp hordei. A quantitative cytological time course analysis revealed that the host cell penetration efficiency in the mutants is intermediate compared with mlo-resistant and Mlo-susceptible genotypes. Ror1 and Ror2 mutants could be differentiated from each other by the same criterion. The spontaneous formation of cell wall appositions in mlo plants, a subcellular structure believed to represent part of the mlo defense, is suppressed in mlo/ror genotypes. In contrast, accumulation of major structural components in the appositions is seemingly unaltered. We conclude that there is a regulatory function for the Ror genes in mlo-specified resistance and propose a model in which the Mlo wild-type allele functions as a negative regulator and the Ror genes act as positive regulators of a non-race-specific resistance response.

(Pubmed less structured abstract)

Discourse segmentation examples

38 of 44 people found the following review helpful:

Move over, Robert Jordan., July 19, 1998

By **A Customer**

This review is from: A Game of Thrones (A Song of Ice and Fire, Book 1) (Mass Market Paperback)

As a fantasy reader of somewhat high standards, I have always had a proclivity for "epic" fantasy. Nothing else really satisfies my desire for an absorbing story. George R.R. Martin has, with this book, taken the field dominated by such giants as Jordan, Williams, and Kay and blown a great big gust of fresh air into it. Not only does this book have the complicated plot and intricate character development that is common to these three talented authors, but it has a certain brutal realism to it. Granted, we're talking about an invented realm, but never before in all the books that I have read has any author taken his portrayal of all the brutality of human nature to this level. Part of what makes Jordan, Williams, and Kay so brilliant is that they write *human* characters, and good and bad are rarely well delineated. What sets Martin apart is his sheer, brutal, mind-numbing honesty. He doesn't pull any punches, and neither do any of his characters. This ! is life, in all its pain and glory. Honor is not as important as we would like it to be, and things do not all go well as long as we wish for it hard enough. Here, there is no destructive force stronger than the power of men. There is no evil greather than that in the hearts of men. And there is no power, once man has decided to destroy, that can stop him. This novel is a masterpiece; beautifully crafted, shockingly realistic, and a joy to read. However, don't expect to come out of reading this with your ideals intact.

Help other customers find the most helpful reviews

Was this review helpful to you?

Report abuse | Permalink

Comment

(5-star Amazon review)

Discourse segmentation examples

41 of 50 people found the following review helpful:

What's left unsaid, February 12, 2004

By **A Customer**

Amazon Verified Purchase (What's this?)

This review is from: A Game of Thrones (A Song of Ice and Fire, Book 1) (Mass Market Paperback)

All of the other excellent reviews of this series are correct. The writing is wonderful. The characters are real. The plot is intricate, fascinating, and never predictable. Et cetera. But none of the reviewers complained about the one thing that has led me to stop reading after plugging through the first two books: This is the darkest, bleakest, most depressing book I have ever read! You must never, ever let yourself bond with a hero, a good, kind, strong, resourceful person who in a 'normal' book would win a gratifying victory at the end of the book. This is because chances are your hero will soon die, most likely brutally. Most (eventually all???) of the good guys die in this book! And everyone is always having to look over his shoulder to see which one of his supposed friends is plotting his death. Innocent children are brutally murdered and their heads put up on pikes. Innocent peasants are slowly hanged, kicking, their eyes bulging out. Their rescuers, instead of pulling off a valiant rescue, are themselves captured and tortured. There are innumerable rapes, including several fairly explicit portrayals of vicious gang rapes of peasant women by invading troops. Every time I finished a reading session I felt depressed. I've never seen so much plague, betrayal, death, and destruction in a novel. It's unrelenting. I don't care how wonderful the writing is. I simply couldn't take it anymore. I want to be uplifted by a book, made to smile and feel vicariously triumphant. I don't want to be beaten down and defeated over and over and over. I had to stop reading.

Help other customers find the most helpful reviews

Was this review helpful to you?

Report abuse | Permalink

Comments (2)

(3-star Amazon review)

Coherence examples

1. Sam brushed his teeth. He got into bed. He felt a certain ennui.
2. Sue was feeling ill. She decided to stay home from work.
3. Sue likes bananas. Jill does not.
4. The senator introduced a new initiative. He hoped to please undecided voters.
5. Linguists like quantifiers. In his lectures, Richard talked only about *every* and *most*.
6. In his lectures, Richard talked only about *every* and *most*. Linguists like quantifiers.

Coherence examples

1. Sam brushed his teeth. then He got into bed. then He felt a certain ennui.
2. Sue was feeling ill. so She decided to stay home from work.
3. Sue likes bananas. but Jill does not.
4. The senator introduced a new initiative. because He hoped to please undecided voters.
5. Linguists like quantifiers. for example In his lectures, Richard talked only about *every* and *most*.
6. In his lectures, Richard talked only about *every* and *most*. in general Linguists like quantifiers.

Coherence examples

1. Sam brushed his teeth. then He got into bed. then He felt a certain ennui.
2. Sue was feeling ill. so She decided to stay home from work.
3. Sue likes bananas. but Jill does not.
4. The senator introduced a new initiative. because He hoped to please undecided voters.
5. Linguists like quantifiers. for example In his lectures, Richard talked only about *every* and *most*.
6. In his lectures, Richard talked only about *every* and *most*. in general Linguists like quantifiers.
7. A: Sue isn't here.
   B: She is feeling ill.
8. A: Where is Bill?
   B: In Bytes Café.
9. A: Pass the cake mix.
   B: Here you go.

(Stone 2002)

Coherence examples

1. Sam brushed his teeth. then He got into bed. then He felt a certain ennui.
2. Sue was feeling ill. so She decided to stay home from work.
3. Sue likes bananas. but Jill does not.
4. The senator introduced a new initiative. because He hoped to please undecided voters.
5. Linguists like quantifiers. for example In his lectures, Richard talked only about *every* and *most*.
6. In his lectures, Richard talked only about *every* and *most*. in general Linguists like quantifiers.
7. A: Sue isn't here.
   B: because She is feeling ill.
8. A: Where is Bill?
   B: answer In Bytes Café.
9. A: Pass the cake mix. (Stone 2002)
   B: fulfillment Here you go.

Discourse segmentation

Clinical Comparison of Full-Field Digital Mammography and Screen-Film Mammography for Detection of Breast Cancer

John M. Lewin[1], Carl J. D'Orsi[2], R. Edward Hendrick[1,3], Lawrence J. Moss[2], Pamela K. Isaacs[1], Andrew Karellas[2] and Gary R. Cutter[4]

[1] University of Colorado Health Sciences Center, 4200 E. 9th Ave., Mail Stop F724, Denver, CO 80262.
[2] University of Massachusetts Medical Center, 55 Lake Ave. N., Worcester, MA 01655.
[3] Northwestern University Medical School, 357 E. Chicago Ave., Chicago, IL 60611.
[4] AMC Cancer Research Center, 1600 Pierce St., Lakewood, CO 80232.

OBJECTIVE. The purpose of this work is to compare full-field digital mammography and screen-film mammography for the detection of breast cancer in a screening population.

SUBJECTS AND METHODS. Full-field digital mammography was performed in addition to screen-film mammography in 6736 examinations of women 40 years old and older presenting for screening mammography at either of two institutions. Two views of each breast were acquired with each technique. The digital and screen-film mammograms were each interpreted independently. In addition to a clinical assessment, each finding was assigned a probability of malignancy for use in receiver operating characteristic analysis. In cases in which the digital and screen-film interpretations differed, a side-by-side analysis was performed to determine the reasons for the discrepancy. With few exceptions, findings detected on either technique were evaluated with additional imaging and, if warranted, biopsy.

RESULTS. Additional evaluation was recommended on at least one technique in 1467 cases. These additional evaluations led to 181 biopsies and the detection of 42 cancers. Nine cancers were detected only on digital mammography, 15 were detected only on screen-film mammography, and 18 were detected on both. The difference in cancer detection is not statistically significant ($p > 0.1$). Digital mammography resulted in fewer recalls than did screenfilm mammography (799 vs 1007, $p < 0.001$). The difference between the receiver operating characteristic curve area for digital (0.74) and screen-film (0.80) mammography was not significant ($p > 0.1$). Reasons for discrepant interpretations of cancer were approximately equally distributed among those relating to lesion conspicuity, lesion appearance, and interpretation.

CONCLUSION. No significant difference in cancer detection was observed between digital mammography and screen-film mammography. Digital mammography resulted in fewer recalls than did screen-film mammography.

Identification of Genes Required for the Function of Non-Race-Specific mlo Resistance to Powdery Mildew in Barley

A. Freialdenhoven, C. Peterhansel, J. Kurth, F. Kreuzaler and P. Schulze-Lefert
Rheinisch-Westfalische Technische Hochschule Aachen, Department of Biology I, Worringer Weg 1, D-52074 Aachen, Germany

Recessive alleles (mlo) of the Mlo locus in barley mediate a broad, non-race-specific resistance reaction to the powdery mildew fungus Erysiphe graminis f sp hordei. A mutational approach was used to identify genes that are required for the function of mlo. Six susceptible M2 individuals were isolated after inoculation with the fungal isolate K1 from chemically mutagenized seed carrying the mlo-5 allele. Susceptibility in each of these individuals is due to monogenic, recessively inherited mutations in loci unlinked to mlo. The mutants identify two unlinked complementation groups, designated Ror1 and Ror2 (required for mlo-specified resistance). Both Ror genes are required for the function of different tested mlo alleles and for mlo function after challenge with different isolates of E. g. f sp hordei. A quantitative cytological time course analysis revealed that the host cell penetration efficiency in the mutants is intermediate compared with mlo-resistant and Mlo-susceptible genotypes. Ror1 and Ror2 mutants could be differentiated from each other by the same criterion. The spontaneous formation of cell wall appositions in mlo plants, a subcellular structure believed to represent part of the mlo defense, is suppressed in mlo/ror genotypes. In contrast, accumulation of major structural components in the appositions is seemingly unaltered. We conclude that there is a regulatory function for the Ror genes in mlo-specified resistance and propose a model in which the Mlo wild-type allele functions as a negative regulator and the Ror genes act as positive regulators of a non-race-specific resistance response.

38 of 44 people found the following review helpful:

Move over, Robert Jordan., July 19, 1998

By **A Customer**

This review is from: **A Game of Thrones (A Song of Ice and Fire, Book 1) (Mass Market Paperback)**

As a fantasy reader of somewhat high standards, I have always had a proclivity for "epic" fantasy. Nothing else really satisfies my desire for an absorbing story. George R.R. Martin has, with this book, taken the field dominated by such giants as Jordan, Williams, and Kay and blown a great big gust of fresh air into it. Not only does this book have the complicated plot and intricate character development that is common to these three talented authors, but it has a certain brutal realism to it. Granted, we're talking about an invented realm, but never before in all the books that I have read has any author taken his portrayal of all the brutality of human nature to this level. Part of what makes Jordan, Williams, and Kay so brilliant is that they write *human* characters, and good and bad are rarely well delineated. What sets Martin apart is his sheer, brutal, mind-numbing honesty. He doesn't pull any punches, and neither do any of his characters. This ! is life, in all its pain and glory. Honor is not as important as we would like it to be, and things do not all go well as long as we wish for it hard enough. Here, there is no destructive force stronger than the power of men. There is no evil greather than that in the hearts of men. And there is no power, once man has decided to destroy, that can stop him. This novel is a masterpiece; beautifully crafted, shockingly realistic, and a joy to read. However, don't expect to come out of reading this with your ideals intact.

Help other customers find the most helpful reviews
Was this review helpful to you?
Report abuse | Permalink
Comment

41 of 50 people found the following review helpful:

What's left unsaid, February 12, 2004

By **A Customer**

Amazon Verified Purchase (What's this?)

This review is from: **A Game of Thrones (A Song of Ice and Fire, Book 1) (Mass Market Paperback)**

All of the other excellent reviews of this series are correct. The writing is wonderful. The characters are real. The plot is intricate, fascinating, and never predictable. Et cetera. But none of the reviewers complained about the one thing that has led me to stop reading after plugging through the first two books: This is the darkest, bleakest, most depressing book I have ever read! You must never, ever let yourself bond with a hero, a good, kind, strong, resourceful person who in a 'normal' book would win a gratifying victory at the end of the book. This is because chances are your hero will soon die, most likely brutally. Most (eventually all???) of the good guys die in this book! And everyone is always having to look over his shoulder to see which one of his supposed friends is plotting his death. Innocent children are brutally murdered and their heads put up on pikes. Innocent peasants are slowly hanged, kicking, their eyes bulging out. Their rescuers, instead of pulling off a valiant rescue, are themselves captured and tortured. There are innumerable rapes, including several fairly explicit portrayals of vicious gang rapes of peasant women by invading troops. Every time I finished a reading session I felt depressed. I've never seen so much plague, betrayal, death, and destruction in a novel. It's unrelenting. I don't care how wonderful the writing is. I simply couldn't take it anymore. I want to be uplifted by a book, made to smile and feel vicariously triumphant. I don't want to be beaten down and defeated over and over and over. I had to stop reading.

Help other customers find the most helpful reviews
Was this review helpful to you?
Report abuse | Permalink
Comments (2)

Discourse segmentation

Hearst's 21-paragraph science news article *Stargazer*

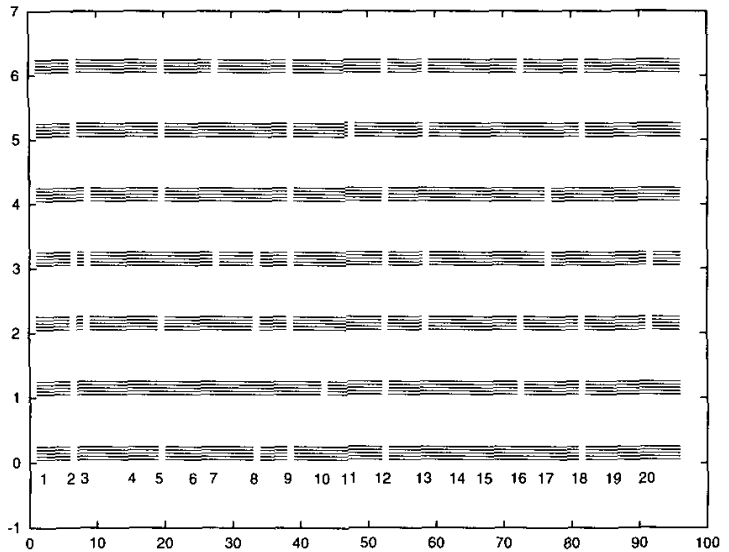
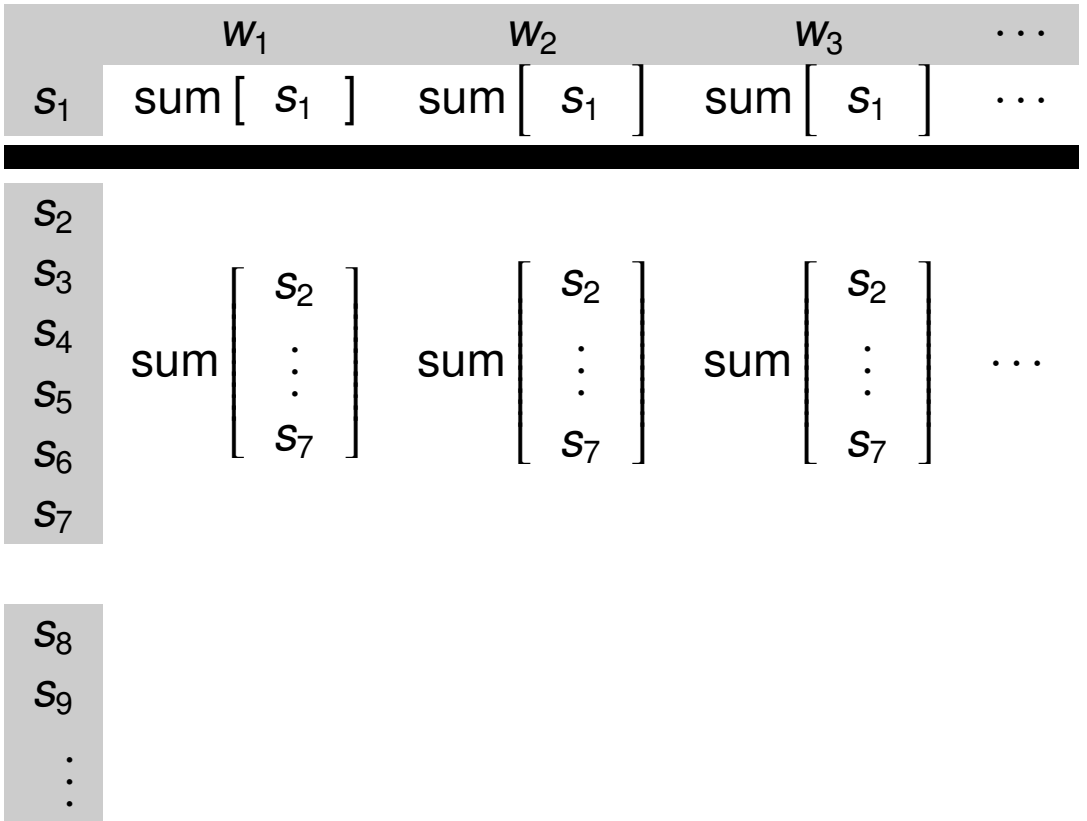


Figure 5
Judgments of seven readers on the *Stargazer* text. Internal numbers indicate location of gaps between paragraphs; *x*-axis indicates token-sequence gap number, *y*-axis indicates judge number, a break in a horizontal line indicates a judge-specified segment break.

1—3	Intro – the search for life in space
4—5	The moon's chemical composition
6—8	How early earth-moon proximity shaped the moon
9—12	How the moon helped life evolve on earth
13	Improbability of the earth-moon system
14—16	Binary/trinary star systems make life unlikely
17—18	The low probability of nonbinary/trinary systems
19—20	Properties of earth's sun that facilitate life
21	Summary

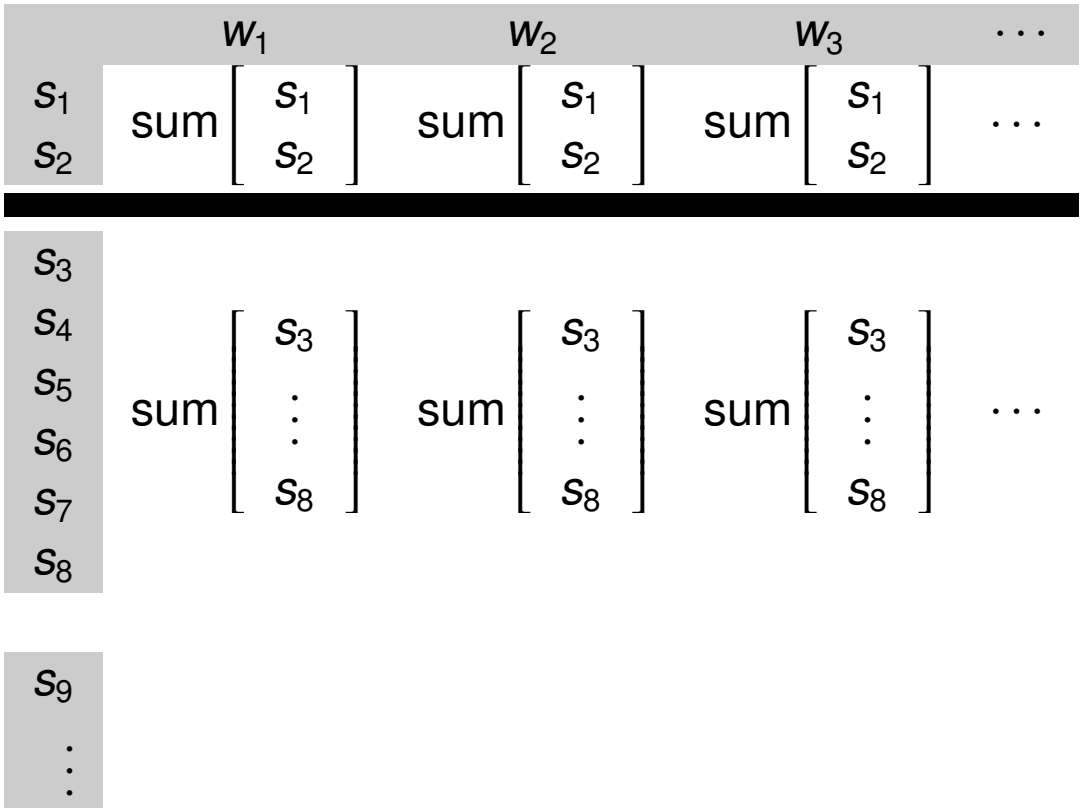
The TextTiling algorithm (Hearst 1994, 1997)



Score this boundary via cosine similarity between the blocks' vectors

Score vector S: $b_{1,2}$

The TextTiling algorithm (Hearst 1994, 1997)



Score this boundary via cosine similarity between the blocks' vectors

Score vector S: $b_{1,2}$ $b_{2,3}$

The TextTiling algorithm (Hearst 1994, 1997)

	w_1	w_2	w_3	$\cdots$
s_1				
s_2	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\cdots$
s_3				
s_4				
s_5				
s_6	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\cdots$
s_7				
s_8				
s_9				
$\vdots$				

Score this boundary via cosine similarity between the blocks' vectors

Score vector S: $b_{1,2}$ $b_{2,3}$ $b_{3,4}$ $\cdots$

The TextTiling algorithm (Hearst 1994, 1997)

	w_1	w_2	w_3	$\cdots$
s_1				
s_2	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$	$\cdots$
s_3				
s_4				
s_5				
s_6	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\text{sum}\begin{bmatrix} s_4 \\ \vdots \\ s_9 \end{bmatrix}$	$\cdots$
s_7				
s_8				
s_9				
$\vdots$				

Score this boundary via cosine similarity between the blocks' vectors

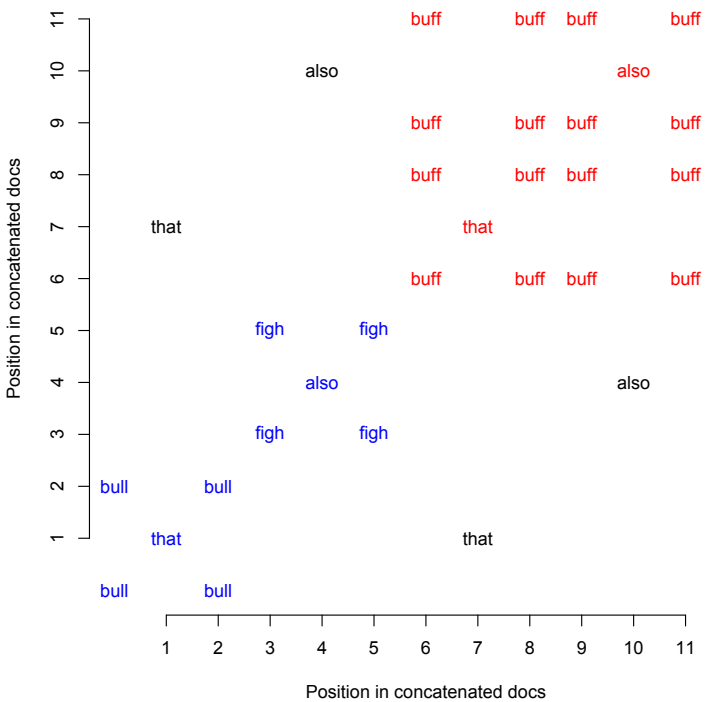
Score vector S: $b_{1,2}$ $b_{2,3}$ $b_{3,4}$ $\cdots$

1. Smooth S using average smoothing over window size a to get $\hat{S}$.
2. Set number of boundaries B as $\mu(\hat{S}) - \frac{\sigma(\hat{S})}{2}$
3. Score each boundary b_i using $(b_{i-1} - b_i) + (b_{i+1} - b_i)$
4. Choose the top B boundaries by these scores.

Dotplotting (Reynar 1994, 1998)

bulldogs	bulldogs	fight	also	fight	buffalo	that	buffalo	buffalo	also	buffalo
1	2	3	4	5	6	7	8	9	10	11

Where word w appears in positions x and y in a single document, add points (x, x), (y, y), (x, y), and (y, x):



Dotplotting (Reynar 1994, 1998)

bulldogs	bulldogs	fight	also	fight	buffalo	that	buffalo	buffalo	also	buffalo
1	2	3	4	5	6	7	8	9	10	11

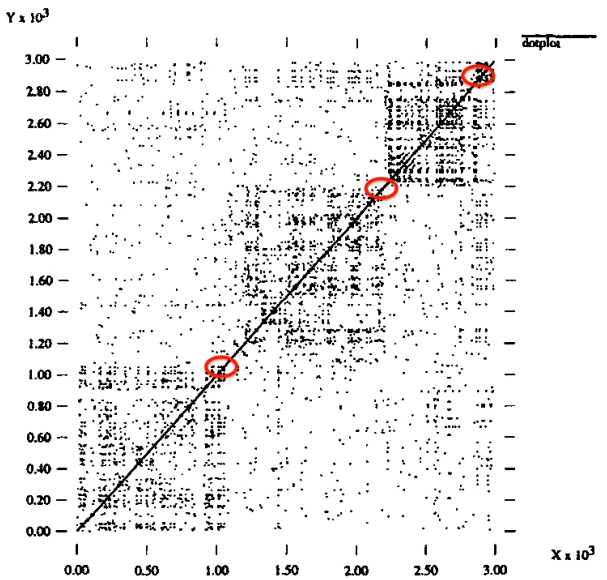
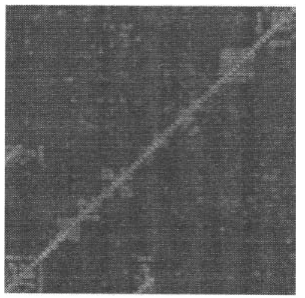


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.* ⬭ = actual doc. boundary

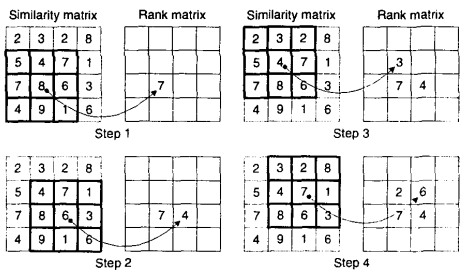
Divisive clustering (Choi 2000)

1. Compare all sentences pairwise for cosine similarity, to create a matrix of similarity values.

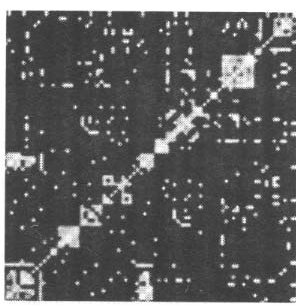


2. For each value s, find the $n \times n$ submatrix N_s with s at its center and replace s with the value

$$\frac{|\{s' \in N_s : s' < s\}|}{n^2}$$



3. Apply something akin to Reynar's algorithm to find the cluster boundaries (which are clearer as a result of the local smoothing



Choi (2000) reports substantial accuracy gains over both TextTiling and dotplotting.

Supervised

1. Label segment boundaries in training and test set.
2. Extract features in training: generally a superset of the features used by unsupervised approaches.
3. Fit a classifier model (NaiveBayes, MaxEnt, SVM, . . .).
4. In testing, apply feature to predict boundaries.

(Manning 1998; Beeferman et al. 1999; Sharp and Chibelushi 2008)

(Slide from Dan Jurafsky.)

Evaluation: WindowDiff (Pevzner and Hearst 2002)

Definition (WindowDiff)

- $b(i, j)$ = the number of boundaries between text positions i and j
- N = the number of sentences

$$\text{WindowDiff}(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} \left(\left| b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}) \right| \neq 0 \right)$$

Return values: 0 = all labels correct; 1 = no labels correct

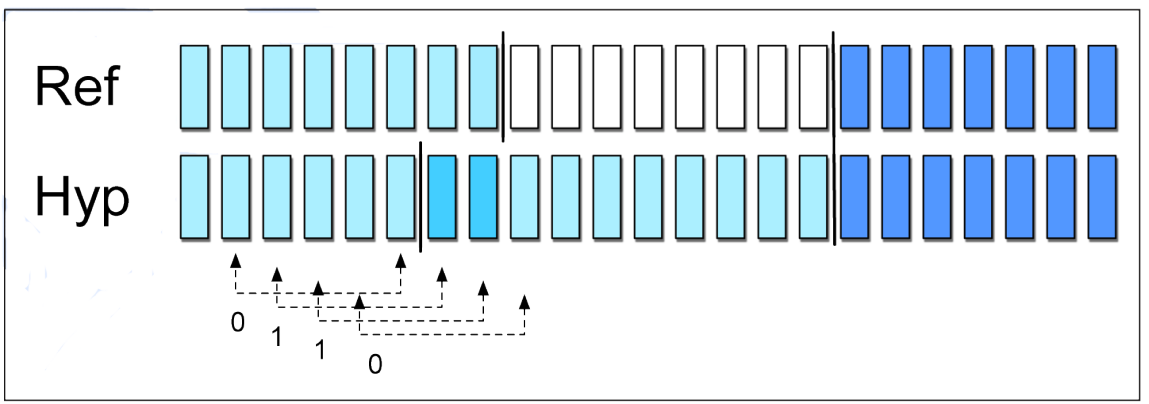


Figure 21.2 The WindowDiff algorithm, showing the moving window sliding over the hypothesis string, and the computation of $|r_i - h_i|$ at four positions. After Pevzner and Hearst (2002).

(Jurafsky and Martin 2009:§21)

Discourse coherence theories

- Halliday and Hasan (1976): Additive, Temporal, Causal, Adversative
- Longacre (1983): Conjoining, Temporal, Implication, Alternation
- Martin (1992): Addition, Temporal, Consequential, Comparison
- Kehler (2002): Result, Explanation, Violated Expectation, Denial of Preventer, Parallel, Contrast (i), Contrast (ii), Exemplification, Generalization, Exception (i), Exception (ii), Elaboration, Occasion (i), Occasion (ii)
- Hobbs (1985): Occasion, Cause, Explanation, Evaluation Background, Exemplification, Elaboration, Parallel, Contrast, Violated Expectation
- Wolf and Gibson (2005): Condition, Violated expectation, Similarity, Contrast, Elaboration, Example, Elaboration, Generalization, Attribution, Temporal Sequence, Same

Rhetorical Structure Theory (RST)

Relations hold between adjacent spans of text: the nucleus and the satellite. Each relation has five fields: constraints on nucleus, constraints on satellite, constraints on nucleus–satellite combination, effect, and locus of effect.

Table 1. *Organization of the relation definitions*

Circumstance	Antithesis and Concession
Solutionhood	Antithesis
Elaboration	Concession
Background	Condition and Otherwise
Enablement and Motivation	Condition
Enablement	Otherwise
Motivation	Interpretation and Evaluation
Evidence and Justify	Interpretation
Evidence	Evaluation
Justify	Restatement and Summary
Relations of Cause	Restatement
Volitional Cause	Summary
Non-Volitional Cause	Other Relations
Volitional Result	Sequence
Non-Volitional Result	Contrast
Purpose	

(Mann and Thompson 1988)

Coherence structures

From Wolf and Gibson (2005)

1. a. Mr. Baker's assistant for inter-American affairs,
   b. Bernard Aronson
2. while maintaining
3. that the Sandinistas had also broken the cease-fire,
4. acknowledged:
5. "It's never very clear who starts what."

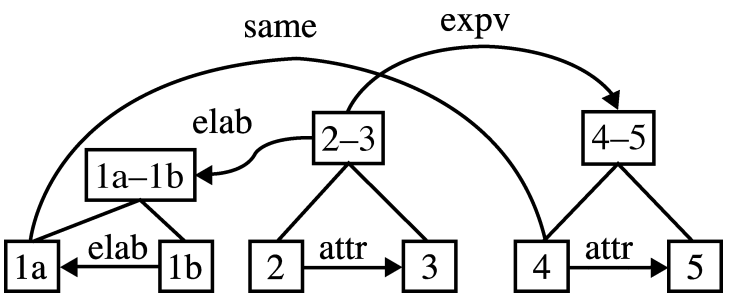


Figure 5
Coherence graph for example (23) with discourse segment 1 split into two segments. *expv* = *violated expectation; elab = elaboration; attr = attribution*.

The Penn Discourse Treebank 2.0 (Webber et al. 2003)

- Large-scale effort to identify the coherence relations that hold between pieces of information in discourse.
- Available from the Linguistic Data Consortium.
- Annotators identified spans of text as the coherence relations. Where the relation was implicit, they picked their own lexical items to fill the role.

Example

[$_{\text{Arg}_1}$ that hung over parts of the factory]
even though
[$_{\text{Arg}_2}$ exhaust fans ventilated the area].

Connectives and their semantics

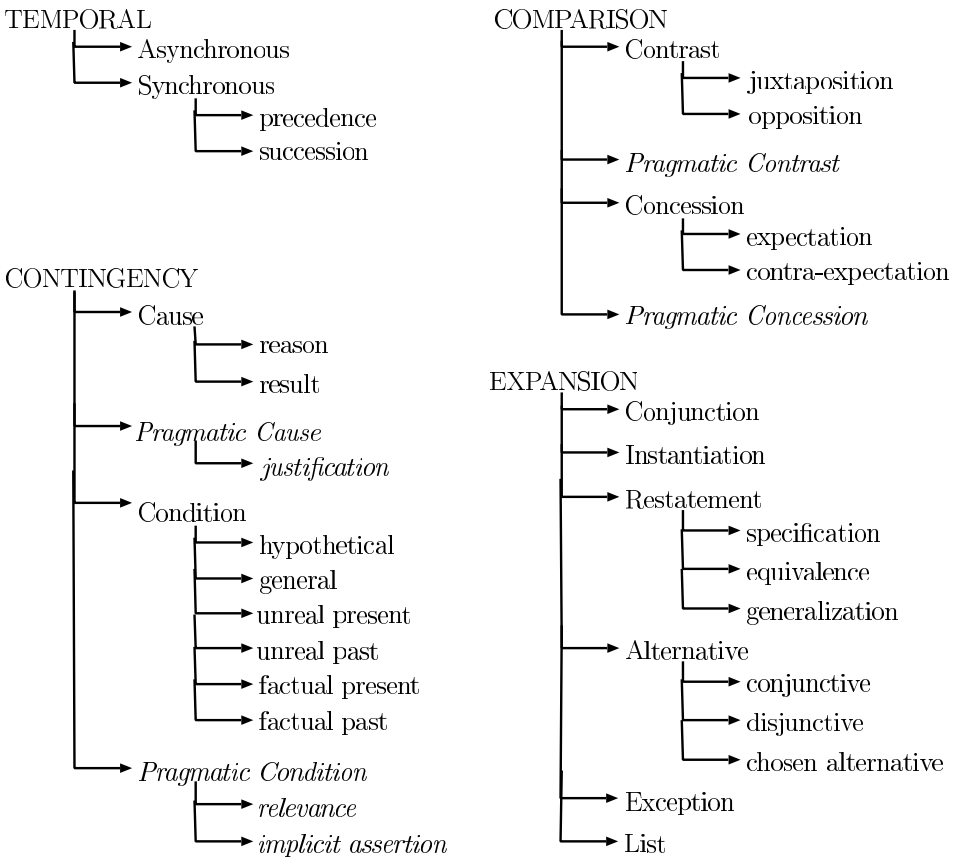


Figure 1: Hierarchy of sense tags

(from Prasad et al. 2008)

Automatically collected labels

Data

- RAW: 41 million sentences (≈1 billion words) from a variety of LDC corpora
- BLIPP: 1.8 million Charniak parsed sentences

Labeling method

1. Extract all sentences matching one of the patterns.
2. Label the connective with the name of the pattern.
3. Treat everything before the connective as Arg1 and everything after it as Arg2.

CONTRAST – 3,881,588 examples [BOS . . . EOS] [BOS But . . . EOS] [BOS . . .] [but . . . EOS] [BOS . . .] [although . . . EOS] [BOS Although . . . ,] [. . . EOS]
CAUSE-EXPLANATION-EVIDENCE — 889,946 examples [BOS . . .] [because . . . EOS] [BOS Because . . . ,] [. . . EOS] [BOS . . . EOS] [BOS Thus, . . . EOS]
CONDITION — 1,203,813 examples [BOS If . . . ,] [. . . EOS] [BOS If . . .] [then . . . EOS] [BOS . . .] [if . . . EOS]
ELABORATION — 1,836,227 examples [BOS . . . EOS] [BOS . . . for example . . . EOS] [BOS . . .] [which . . . ,]
NO-RELATION-SAME-TEXT — 1,000,000 examples Randomly extract two sentences that are more than 3 sentences apart in a given text.
NO-RELATION-DIFFERENT-TEXTS — 1,000,000 examples Randomly extract two sentences from two different documents.

Table 2: Patterns used to automatically construct a corpus of text span pairs labeled with discourse relations.

Data and tools

- Penn Discourse Treebank 2.0
  - LDC: `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T05`
  - Project page: `http://www.seas.upenn.edu/~pdtb/`
  - Python tools/code: `http://compprag.christopherpotts.net/pdtb.html`
- Rhetorical Structure Theory
  - LDC: `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002T07`
  - Project page: `http://www.sfu.ca/rst/`

Prospects

Text segmentation

Seems to have fallen out of fashion, but obviously important to many kinds of information extraction — probably awaiting a breakthrough idea.

Discourse coherence

On the rise in linguistics but perhaps not in NLP. Essential to all aspects of NLU, though, so a breakthrough would probably have widespread influence.