Discourse Processing
A Tutorial

Manfred Stede
Applied Computational Linguistics
EB Cognitive Science
University of Potsdam
Germany

- Most of the material is based on

  Manfred Stede: Discourse Processing. Morgan & Claypool 2011.

„Discourse Processing“

- Where does it start?
- Treat your text document
  - not as a bag of words,
  - not as a bag of sentences,
  - but as a linear sequence of connected sentences, adding hierarchical structure where appropriate

Discourse-sensitivity in NLP Tasks

- Example: information extraction (Wikipedia)

Angela Dorothea Merkel (born 17 July 1954) is a German politician who has been the Chancellor of Germany since 2005, and the Leader of the Christian Democratic Union (CDU) since 2000. She is the first woman to hold either office.

Discourse-sensitivity in NLP Tasks

- Example: summarization

When you know the global structure of your text document, you can make sure to cover relevant portions

(examples will follow...)

Discourse-sensitivity in NLP Tasks

- Example: opinion mining

(...) *We didn‘t like the village very much and won‘t come back because there are only few things to see, and moreover the place is quite dirty. (...)*

Goal for today

- Provide an overview of the major problems of discourse processing on text documents, and on the central ideas for tackling them

Tutorial overview

- Part 1: „Large“ discourse units
  - Genre-specific structure
  - Topics
- Part 2: Coreference
- Part 3: „Small“ discourse units
  - Local coherence
  - Coherence-relational text structure
- Exploring interconnections

Genre: Court decision

BUNDESGERICHTSHOF

IM NAMEN DES VOLKES

URTEIL

VI ZR 101/06

Verkündet am:
27. März 2007
Holmes,
Justizangestellte
als Urkundsbeamtin
der Geschäftsstelle

in dem Rechtsstreit

Nachschlagewerk: ja
BGHZ: nein
BGHR: ja

BGB § 823 Ah, G, § 1004, StGB § 185, TMG § 10

Ein Unterlassungsanspruch wegen eines in ein Meinungsforum im Internet eingestellten ehrverletzenden Beitrags kann auch dann gegen den Betreiber des Forums gegeben sein, wenn dem Verletzten die Identität des Autors bekannt ist.

BUNDESGERICHTSHOF

IM NAMEN DES VOLKES

URTEIL

VI ZR 101/06

Verkündet am:
27. März 2007
Holmes,
Justizangestellte
als Urkundsbeamtin
der Geschäftsstelle

in dem Rechtsstreit

Nachschlagewerk: ja
BGHZ: nein
BGHR: ja

BGB § 823 Ah, G, § 1004, StGB § 185, TMG § 10

Ein Unterlassungsanspruch wegen eines in ein Meinungsforum im Internet eingestellten ehrverletzenden Beitrags kann auch dann gegen den Betreiber des Forums gegeben sein, wenn dem Verletzten die Identität des Autors bekannt ist.

Genre: Scientific paper
(Teufel et al. 09)

Synthesis of pyrazole and pyrimidine Troeger's base–analogues

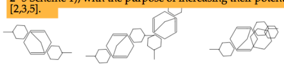
1 PERKIN

Rodrigo Abonia, Andrea Albernez, Hector Larrabondo, Jairo Quiroga, Braulio Isuasty, Henry Isuasty, Angelina Hormaza Adolfo Sanchez, and Manuel Nogueras

Troeger's–base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3–alkyl–5–amino–1arylyrazoles and 6–aminopyrimidin–4(3H)–ones with formaldehyde under mild conditions (i.e. in ethanol at 50C in the presence of catalytic amounts of acetic acid. Two key intermediates were isolated from the reaction mixtures, which helped us to suggest a sequence of steps for the formation of the Troeger's bases obtained. The structures of the products were assigned by 1H and 13 CNMR, mass spectra and elemental analysis and confirmed by X–ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base 1 was obtained more than a century ago from the raction of p–toluidine and formaldehyde [11], recently the study of these compounds has gained importance due to their potential applications. They possess a relatively rigid chiral structure which makes them suitable for the development of possible synthetic enzyme and artificial receptor systems [2], chelating and biomimetic systems [3] and transition metal complexes for regio–and stereoselective catalytic reac–tions [4]. For these reasons, numerous Troeger's–base derivates have been prepared bearing different types of substituents and structures (i.e. 2–5 Scheme 1), with the purpose of increasing their potential applications [2,3,5].



Scheme 1 The original Troeger's–base 1 and some interesting deri–vatives and analogues.

However, some of the above methodologies possess tedious work–up procedures or include relatively strong reaction conditions, such as treatment of the starting materials for several hours with an ethanolic solution of conc. hydrochloric acid or TFA solution, with poor to moderate yields, as is the case for analogues 4 and 5 [5].

Considering these potential applications, we now report a simple synthetic method for the preparation of 5,12–dialkyl–3,10–diaryl–1,3,4,8,10,11–hexa–azetetracyclo[6.6.1.0 2,6 .0 9,13] pentadeca–2(6),4,9(13),11–tetraenes 8a–e and 4,12–dimethoxy–1,3,5,9,11,13–hezaaatetrctyclo[7.7.1.0 2,7.010,15] heptadeca2(7),3,10(15)m11–tetraene–6m14–diones 10a,b based on thereaction of 3–alkyl–5–amino–1–arylpyrazoles 6 and 6–aminopyrimin–4(3H)–ones 9 with formaldehyde in ethanol and catalytic amounts of acetic acid. Compounds 8 and 10 are new Troegers base analogues bearing heterocyclic rings instead of the usual phenyl rings in their aromatic parts.

Results and discussion

In an attempt to prepare the benzotriazolyl derivative 7a, which could be used as in intermediate in the synthesis of new hydroquinolines of interest, [6], a mixture of 5–amino–3–methy–1–phenylpyrazole 6a,formaldehyde and benz,otri–azole in 10 ml of ethanol , with catalytic amounts of acetic acid, weas heated at 50C for 5 minutes. A solid precipidated from the solution while it was still hot. However, no consumption of benzotriazole was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazoole, as shown in Schema 12. On the basis of NMR and mass spectra and X–ray crystallographic analysis we established that the structure is 5,12–diakyl–3 10–diaryl–1,3,4,8,10,11–hexa azetatetracyclo[6.6.1.0 2,6 .0 9,13]penta–deca–2.6[4],9[13]11tetraene 8, a new pentacyclic Troeger's base analogue.

Co_Gro	Other	Aim	Gap/Weak	Own_Mthd	Own_Res	Own_Conc

Teufel et al. 09: Features

- Absolute location
  - pos. of sentence in doc
- Explicit structure
  - pos. of sentence in sct
  - pos. of sentence in para
  - type of headline of sct
- Sentence length
  - >12
- Content features
  - sentence contains words from title or headlines
  - sentence contains tf/idf-prominent terms
- Verb syntax
  - voice of first verb in sentence
  - tense of first verb in sentence
  - Is first verb modified by aux
- Citations
  - is citation present
  - self or other
  - positions
- History
  - most likely previous zone
- Meta-discourse
  - type of formulaic expr (28)
  - type of agent (10)
  - verb class and negation presence (28)

Teufel et al. 09: Results

- Chemistry / CompLing
- 110pp guidelines
- 30 papers annotated
- Human Kappa = .71 / .65
- Automatic Kappa = .41 / .39
- Zone classification f-measure: 26%-86%

Genre: Newspaper Editorial

Should Berlin apply for the 2016 Olympics?

[2] Hamburg has long understood: [3] Olympic games are worth a lot of gold. [4] Those who draw the Olympics into their city are winners in the world-wide competition for attention. [5] That's why Berlin must not miss the opportunity for the 2016 games. [6] The capital must grab the baton from Leipzig and apply to be the venue. [7] Barcelona has shown that the olympic effect is invaluable. [8] With the 1992 games the city has re-invented itself -- and makes profit up to today: [9] The number of overnight stays has doubled, the economy is still profiting. [10] When Berlin now runs again as applicant, we show the world that we're better now than we were once. [11] After all, today the city offers what a candidate needs: [12] big-city flair, hotel beds, infrastructure. [13] The sports venues planned for 2000, such as Velodrom and Max-Schmeling-Halle, exist, the olympic stadium is in mint condition, the Anschütz arena is nearing completion. [14] Just by re-applying, Berlin would already modernize itself and improve its international profile. [15] Public and private sponsoring money would pour in, millions would follow from IOC. [16] And even if a European city turns out to be the venue for 2012: [17] One has to flex one's muscles in order to win the games, if necessary with the third instead of the second application. [18] Berlin to the starting block: [19] On your mark, ready, go!

Should Berlin apply for the 2016 Olympics?

[2] Hamburg has long understood: [3] Olympic games are worth a lot of gold. [4] Those who draw the Olympics into their city are winners in the world-wide competition for attention. [5] That's why Berlin must not miss the opportunity for the 2016 games. [6] The capital must grab the baton from Leipzig and apply to be the venue. [7] Barcelona has shown that the olympic effect is invaluable. [8] With the 1992 games the city has re-invented itself -- and makes profit up to today: [9] The number of overnight stays has doubled, the economy is still profit-ing. [10] When Berlin now runs again as applicant, we show the world that we're better now than we were once. [11] After all, today the city offers what a candidate needs: [12] big-city flair, hotel beds, infrastructure. [13] The sports venues planned for 2000, such as Velodrom and Max-Schmeling-Halle, exist, the olympic stadium is in mint condition, the Anschütz arena is nearing completion. [14] Just by re-applying, Berlin would already modernize itself and improve its international profile. [15] Public and private sponsoring money would pour in, millions would follow from IOC. [16] And even if a European city turns out to be the venue for 2012: [17] One has to flex one's muscles in order to win the games, if necessary with the third instead of the second application. [18] Berlin to the starting block: [19] On your mark, ready, go!

Content zones of *Pro&Contra* Editorial

- (Introduction, exposition of the problem)
- Central thesis of author
- Argument, pro-author
- (Counter-argument, contra-author)
- (Refutation of counter-argument)
- (Background information)
- (Final statement, rhetorical ending)

(optional zone)
linear position fixed

Genre: Film review

The Draughtsman's Contract

by James Mackenzie

James Mackenzie is currently finishing a Bachelor of Arts at Adelaide University, majoring in philosophy.

The Draughtsman's Contract (1982 UK 108 mins)

Source: CAC/NLA **Prod Co**: BFI/Channel 4 **Prod**: David Payne **Dir, Scr**: Peter Greenaway **Ph**: Curtis Clark **Ed**: John Wilson **Art Dir**: Bob Ringwood **Mus**: Michael Nyman

Cast: Anthony Higgins, Janet Suzman, Anne Louise Lambert, Hugh Fraser, Suzanne Crowley, Neil Cunningham

In 1981, Peter Greenaway spent a warm summer drawing his house on the Welsh Border. So that the house was in the same light every time he sketched it, eight set views were drawn at eight set times of the day. The carefulness of the plan balanced the chaos of its enactment: cows, neighbours and children were equally constant, if pleasurable, interruptions. Anybody who has seen *The Draughtsman's Contract* may find this scenario familiar. It was the basis of the film. But the relevance of any biographical detail stops there.

Like all of Greenaway's work, *The Draughtsman's Contract* addresses issues of representation: between mediums (drawn and photographic representation); between art and nature (the inbreeding between landscape art and ornamental garden design); and the value of classical naturalism in art.

Genre structure – Content zones

(Bieler et al. 07)

Describe { Plot, Character, Specific, General, Content }

Comment { Plot, Actors+characters, Specific, General, Overall }

Describe+Comment { Plot, Actors+characters, Specific, General, Content }

Quote

Background

Interpretation

Figure A1. Functional zones

Structural elements { Tagline, Structure, Off-topic }

Information about the film { Title, Title+year, Runtime, Country+year, Director, Genre, Audience-restriction, Cast, Credits, Show-Loc+date, Misc-Movie-Info }

Information about the review { Source, Author, Author-Bio, Place, Date, Legal-Notice, Misc-Review-Info, Rating }

Figure A2. Formal zones

Describe vs. Comment Paragraphs

Architect Stourley Kracklite (Brian Dennehy) arrives in Rome, where an exhibition of the works of the 18th-century architect Etienne-Louis Boullée is being mounted under Kracklite's supervision. The city - or something - doesn't sit with him; upon arrival, he begins complaining of stomach pains. Cancer? Kracklite is sure of it. Or not: It could be that his wife Louisa (Chloe Webb), with whom he is traveling (and who is pregnant with his child), is poisoning him, a revenge for his self-absorption. She may be further motivated in this by the affair she has taken up with Caspasian Speckler (Lambert Wilson), another architect involved with the exhibition.

Greenaway has an eye for composition, and in *The Belly of an Architect* many formal arrangements stand out for their beauty. Dennehy, always engaging, is slyly illegible in the central role, a stroke of luck maybe for the director, who has shown himself to be disinclined to bother much with actors. But the relentless condescension and self-congratulation with which Greenaway conducts this very private amusement is grotesque. He fosters the worst imaginable relationship with his audience: showing off while condemning those not enlightened enough to cherish his preening. In Kracklite, Greenaway has created a self-obsessed, boorish non-hero on whom to hang his obscurantist ramblings, and his indifference to his audience is so great that he expects us to relish it. Who's the asshole here?

Describe vs. Comment: bag-of-terms
(Bieler et al. 07)

- Supervised learning (SVM)
- Development set: 100 German reviews from 5 internet sites
- Training/test set: 112 reviews, 60 of which come from three „new“ sites
  - Training set (66%)
  - Test set (33%)
- Features = character 5-grams, weighted with TF/IDF against reference corpus

Zone type	*Comment*	*Describe*
Precision	81.6	76.8
Recall	79.7	79.0

Classifying zones: English texts
(Taboada et al. 09)

- Manual annotation
  – Three annotators to check agreement

Classes	2-rater kappa	3-rater kappa
Describe/Comment/Describe+Comment/Formal	.82	.73
Describe/Comment/Formal	.92	.84
Describe/Comment/Describe+Comment	.68	.54
Describe/Comment	.84	.69

Table B1. Kappa values for stage annotations

  – Then, one annotator annotated 100 texts from RottenTomatoes.com
  – 83.000 words in 1.500 paragraphs

Stage	Count
Describe	347
Comment	237
Describe+Comment	237
Background	51
Interpretation	22
Quote	2
Formal	646

Table 1. Stages in 100 text RT corpus

Automatic classification: Features

- Character 5-grams that appear at least 4 times in the corpus
- Genre-based
  1. From Biber (88)
     - 1, 2, 3 person pronouns; demonstrative pronouns
     - Place and time adverbials
     - Intensifiers
     - Modals
  2. Connectives that indicate contrast, comparison, causation, evidence, condition
  3. List of 500 adjectives classified in terms of Appraisal (Martin and White 2005)
     - Appreciation, Judgment or Affect
  4. Text statistics
     - Average length of words and sentences; position of paragraphs in the text

Classifier performance

2, 3, and 4 classes
- Comment, Describe
- Comment, Describe, Formal
- Comment, Describe, Comment+Describe, Formal

10-fold cross validation

Classifier	Comment			Describe			Formal			Desc+Comm			Overall Accuracy
	P	R	F	P	R	F	P	R	F	P	R	F	
2-class-5-gram-Bayes	.66	**.79**	.72	.70	.55	.62	-	-	-	-	-	-	68.0
2-class-5-gram-SVM	.53	.63	.64	.68	.69	**.69**	-	-	-	-	-	-	66.8
2-class-genre-Bayes	.66	.75	.70	.67	.57	.61	-	-	-	-	-	-	66.2
2-class-genre-SVM	**.71**	.76	**.74**	**.71**	.65	.68	-	-	-	-	-	-	**71.1**
3-class-5-gram-Bayes	.69	.49	.57	.66	**.78**	**.71**	**.92**	**.97**	**.95**	-	-	-	**78.1**
3-class-5-gram-SVM	.64	.63	.63	**.68**	.65	.65	.91	**.97**	.94	-	-	-	77.2
3-class-genre-Bayes	**.68**	.68	.66	.67	.46	.55	.84	.96	.90	-	-	-	74.0
3-class-genre-SVM	.66	**.71**	**.68**	.67	.56	.61	.90	.94	.92	-	-	-	76.8
4-class-5-gram-Bayes	**.46**	.35	.38	**.69**	.47	.56	**.92**	**.97**	**.95**	**.42**	**.64**	**.51**	69.0
4-class-5-gram-SVM	.43	**.41**	.44	.59	.62	.60	.91	**.97**	.94	.45	.41	.42	**69.6**
4-class-genre-Bayes	.38	.31	.34	.66	.30	.41	.86	**.97**	.90	.33	.60	.42	62.3
4-class-genre-SVM	.46	.32	.38	.53	**.82**	.65	.87	.94	.90	.26	.03	.06	67.4

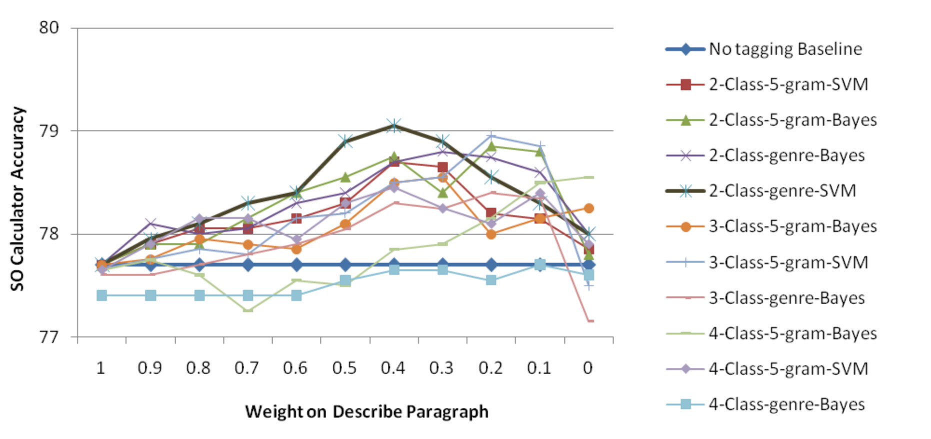
Table 2. Stage identification performance of various categorical classifiers

Content zones and Sentiment detection

- Idea: Use zone-classifier to improve the performance of text-level sentiment classification by *SO-CAL* (Taboada et al. 11) on movie reviews
  - Disregard description altogether (weight of 0)
  - Give it a lower weight than to comment
- Evaluation performed on the *Polarity Dataset*, a collection of 2,000 movie reviews (Pang/Lee 04)
  - Run the zone classifiers to label paragraphs
  - Run SO-CAL on the texts, with different weights assigned to each type of paragraph

SO-CAL and weighing paragraphs

- On the Polarity Dataset, 2,000 movie reviews (Pang and Lee 2004)



SO-CAL and weighing paragraphs

- Most classifiers improve performance over the 77.7% baseline
- Using manual zone annotations (the original 100 texts used to build the classifiers), performance is boosted by 12%
- Precision of the classifier is key
  - Need to identify Describe paragraphs accurately
  - In lieu of that, weighting is a better strategy than removing all Describe paragraphs

© Manfred Stede / NAACL Tutorial 2013

Wrap-up: Definitions

- Content zone

  A continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text.

- Genre

  A class of texts that fulfill a common function, are being used for a common communicative purpose, and are potentially subject to conventions on various levels of description:
  - length
  - layout (headlines, pictures, tables, diagrams, enumerations, ...)
  - lexical and syntactic features
  - internal organization
    - Which zones are obligatory, which are optional
    - Constraints and preferences on zone order

Tutorial overview

- Part 1: „Large“ discourse units
  - Genre-specific structure
  - Topics
- Part 2: Coreference
- Part 3: „Small“ discourse units
  - Local coherence
  - Coherence-relational text structure
- Exploring interconnections

[1.7] A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there. [1.8]"I don't see him—it's a catastrophe," Gaedi said. [1.9]"God gives, God takes." [1.10] Chris Rolling, an American missionary and aid worker, tried to extricate a girl named Jacqueline from a collapsed school using nothing more than a hammer. [1.11] He urged her to be calm and pray, and as night fell he promised that he would return with help. [1.12] When he came back the next morning, Jacqueline was dead. [1.13]"The bodies stopped bothering me after a while, but I think what I will always carry with me is the conversation I had with Jacqueline before I left her," Rolling wrote afterward on his blog. [1.14]"How could I leave someone who was dying, trapped in a building! ...[1.15] She seemed so brave when I left! [1.16] I told her I was going to get help, but I didn't tell her I would be gone until morning. [1.17] I think this is going to trouble me for a long time." [1.18] Dozens of readers wrote to comfort Rolling with the view that his story was evidence of divine wisdom and mercy.

[1.7] A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there. [1.8]"I don't see him—it's a catastrophe," Gaedi said. [1.9]"God gives, God takes." [1.10] Chris Rolling, an American missionary and aid worker, tried to extricate a girl named Jacqueline from a collapsed school using nothing more than a hammer. [1.11] He urged her to be calm and pray, and as night fell he promised that he would return with help. [1.12] When he came back the next morning, Jacqueline was dead. [1.13]"The bodies stopped bothering me after a while, but I think what I will always carry with me is the conversation I had with Jacqueline before I left her," Rolling wrote afterward on his blog. [1.14]"How could I leave someone who was dying, trapped in a building! ...[1.15] She seemed so brave when I left! [1.16] I told her I was going to get help, but I didn't tell her I would be gone until morning. [1.17] I think this is going to trouble me for a long time." [1.18] Dozens of readers wrote to comfort Rolling with the view that his story was evidence of divine wisdom and mercy.

Definition

- Topic-induced text structure

A sequence of non-overlapping text segments that completely covers the text, i.e., a partitioning. Each unit consists of one or more sentences that address a common topic.

Surface Cues

- Paragraph breaks
  - *Prima facie* a good cue, but is to be handled with care
  - Cautious approach: If a text has many paragraph breaks, topic breaks are unlikely to occur *within* paragraphs
- Connectives
  - can indicate continuity: *also, then, so, ...*
  - can indicate discontinuity: *but, still, besides, ...*
- Pronouns
  - often indicate continuity
- Syntax: information structure

- => altogether not very reliable (or difficult to compute)

Content Words (1): Lexical chains

- Intuition: topic continuity means that words in contiguous sentences are related
- Steps:
  - (a) Compute relations between individual words
  - (b) Build *chains* from those relations
  - (c) Induce boundaries from the topology of chains

(a) Establish Lexical Relations

- Which words are to be considered as candidates?
  - Part of speech: usually nouns and NEs
  - Distance:
    - Fixed: a few sentences
    - Flexible: Hirst/St. Onge (98) allow longer distance for stronger relations

(a) Establish Lexical Relations

- Measurement
  - WordNet path length (e.g. *WordNet Connect*)
    Problem: need to do word sense disambiguation, or not (see Silber/McCoy 02)
  - Distributional similarity

(b) Build chains

- Move through text, deciding for each word to
  - start a new chain
  - connect to an existing chain
    Caveat: do not compare to the last item only! (Morris/Hirst 91: *cow – sheep – wool – scarf – boots – hat – snow*)
- Need weights and thresholds for chain length, chain density, chain distance

(c) Induce topic boundaries

- Compute boundary strength for each sentence break
  - Number of chains ending, beginning, crossing
- Accept the top *n* boundary candidates

- Some implementations available, e.g. *LCSeg* www.cs.columbia.edu/nlp/tools.cgi

Content Words (2): Vocabulary shifts

- Can we find those boundaries without using lexical resources?

- Straightforward idea:
  - Word repetition
  - Introduction of new words

Reynar 94: *dot plotting*

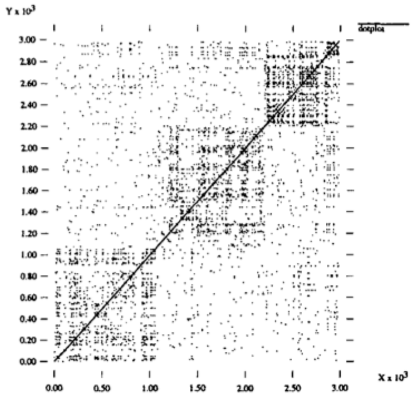


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

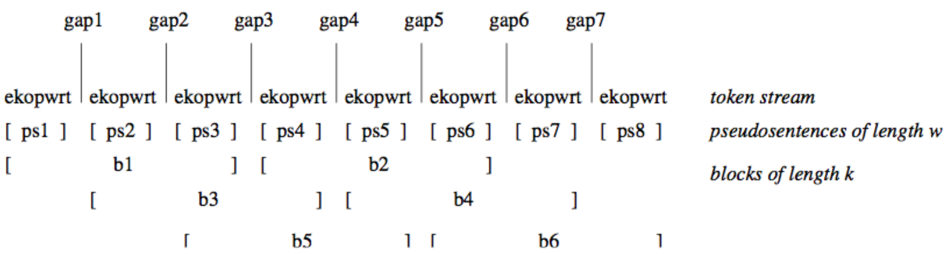
- Closed-class words removed
- Forms of *to be, to have* removed
- Lemmatization applied

Word appears at pos x and y
=> dot at (x,x) (y,y) (x,y) (y,x)

At least for synthetic data, this works at least as good as lexical chaining (e.g., Stokes et al. 04)

- Skorochod'ko (72):
  - divide the text into sentences,
  - count the overlap in content words between neighboring sentences,
  - postulate topic boundaries on the basis of the overlap count.
- Various implementations, including *Text Tiling* (Hearst 97)
  - work on *expository* text (as opposed to narrative, descriptive, argumentative, instructive)

Text tiling (Hearst 1997)



- Recommendation: w=20, k=6
- At every pseudosentence boundary, compute similarity between blocks meeting there

Text tiling: similarity

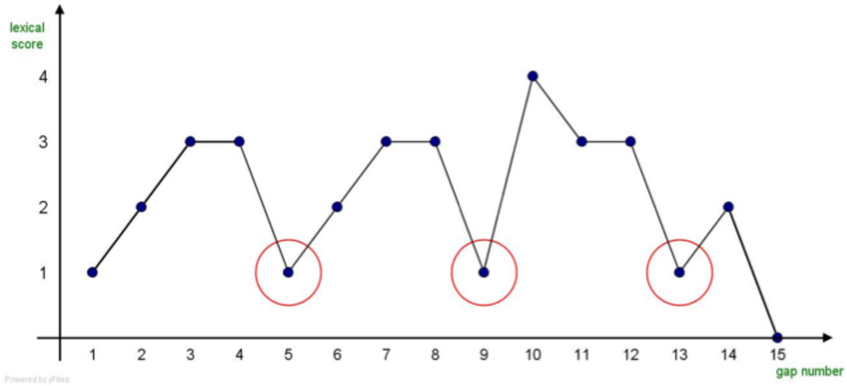
- Hearst 97: Similarity via cosine metric

$$score(gap_i) = \frac{\sum_t w_{t,b1} w_{t,b2}}{\sqrt{\sum_t w_{t,b1}^2 \sum_t w_{t,b2}^2}}$$

- Choi 00: Need to consider the distribution of words across the whole text – not just locally
- Dias et al. 07: word relevance also depends on *inverse document frequency,* adapted to sentences:

$$tf.isf(word) = \frac{stf(word, s)}{|s|} * \ln \frac{Ns}{sf(word)}$$

Text tiling: boundaries

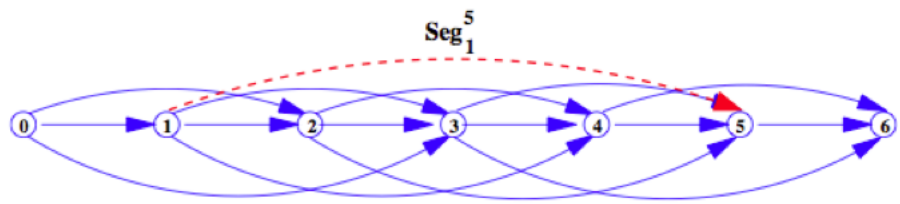


Hearst 97: compute depth score for each gap, then accept the top n gaps
depth(gap_i) = $(y_{i-1} - y_i) + (y_{i+1} - y_i)$

Dias et al. 07: compare steepness of the function at valleys

Other implementations

- Shortest-path problem (Misra et al. 09)



Other options

- Combine word distribution analysis with surface cues, e.g. Beeferman et al. (99), Galley et al. (03)
- Current line of work: compute *hidden topics*, using latent Dirichlet allocation (LDA), e.g. Eisenstein (09)

Tutorial overview

- Part 1: „Large“ discourse units
  - Genre-specific structure
  - Topics
- Part 2: Coreference
- Part 3: „Small“ discourse units
  - Local coherence
  - Coherence-relational text structure
- Exploring interconnections

Ref. exprs. / Mentions / Markables

[1.7] A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there. [1.8]"I don't see him—it's a catastrophe," Gaedi said. [1.9]"God gives, God takes." [1.10] Chris Rolling, an American missionary and aid worker, tried to extricate a girl named Jacqueline from a collapsed school using nothing more than a hammer. [1.11] He urged her to be calm and pray, and as night fell he promised that he would return with help. [1.12] When he came back the next morning, Jacqueline was dead. [1.13]"The bodies stopped bothering me after a while, but I think what I will always carry with me is the conversation I had with Jacqueline before I left her," Rolling wrote afterward on his blog. [1.14]"How could I leave someone who was dying, trapped in a building! ...[1.15] She seemed so brave when I left! [1.16] I told her I was going to get help, but I didn't tell her I would be gone until morning. [1.17] I think this is going to trouble me for a long time." [1.18] Dozens of readers wrote to comfort Rolling with the view that his story was evidence of divine wisdom and mercy.

Referential Chains

[1.7] A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there. [1.8]"I don't see him—it's a catastrophe," Gaedi said. [1.9]"God gives, God takes." [1.10] Chris Rolling, an American missionary and aid worker, tried to extricate a girl named Jacqueline from a collapsed school using nothing more than a hammer. [1.11] He urged her to be calm and pray, and as night fell he promised that he would return with help. [1.12] When he came back the next morning, Jacqueline was dead. [1.13]"The bodies stopped bothering me after a while, but I think what I will always carry with me is the conversation I had with Jacqueline before I left her," Rolling wrote afterward on his blog. [1.14]"How could I leave someone who was dying, trapped in a building! ...[1.15] She seemed so brave when I left! [1.16] I told her I was going to get help, but I didn't tell her I would be gone until morning. [1.17] I think this is going to trouble me for a long time." [1.18] Dozens of readers wrote to comfort Rolling with the view that his story was evidence of divine wisdom and mercy.

Two tasks

- *Anaphor* = referring expression that cannot be resolved without finding its antecedent in the preceding context

- Anaphora resolution:
  - find an antecedent for each anaphor in a text.
  - 'Anaphora' is an irreflexive, non-symmetrical relation.
- Co-/reference resolution:
  - partition the set of mentions of discourse referents in a text into classes (chains).
  - Since referents are identical, 'coreference' is an equivalence relation (reflexive, symmetrical, transitive)

- Co-referent and anaphoric:

  Jim forgot his umbrella. He had to return to his house.

- Co-referent but not anaphoric:

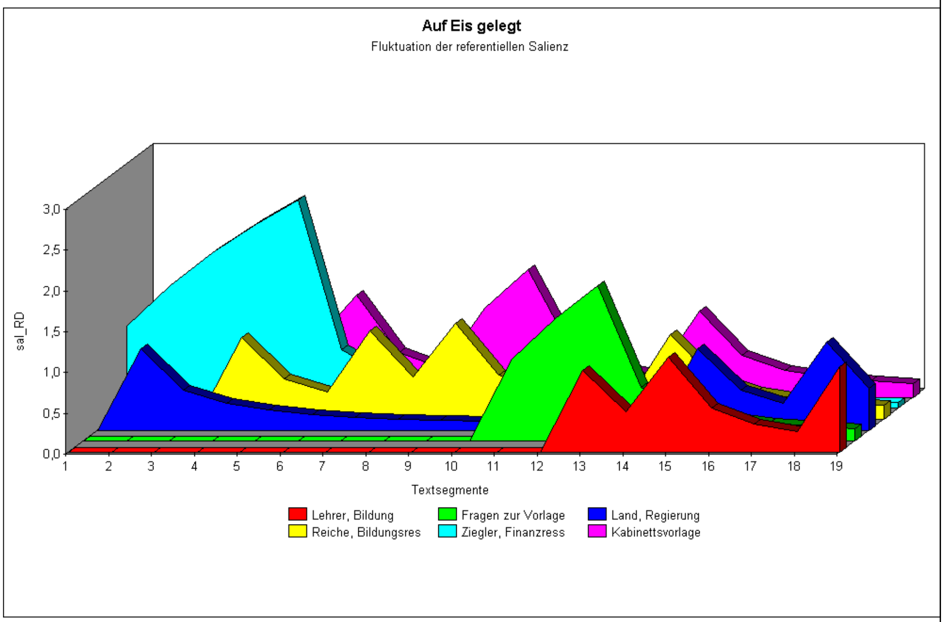
  Apple, Inc. announced a record surplus for the third quarter of the year. Overall, it has been a very successful year for Apple.

- Not co-referent but anaphoric:

  Potsdam University enrolled 1500 new students this year. Therefore, the president is very enthusiastic.

AUF EIS GELEGT

Dagmar Ziegler sitzt in der Schuldenfalle. Auf Grund der dramatischen Kassenlage in Brandenburg hat sie jetzt eine seit mehr als einem Jahr erarbeitete Kabinettsvorlage überraschend auf Eis gelegt und vorgeschlagen, erst 2003 darüber zu entscheiden. Überraschend, weil das Finanz- und das Bildungsressort das Lehrerpersonalkonzept gemeinsam entwickelt hatten. Der Rückzieher der Finanzministerin ist aber verständlich. Es dürfte derzeit schwer zu vermitteln sein, weshalb ein Ressort pauschal von künftigen Einsparungen ausgenommen werden soll - auf Kosten der anderen. Reiches Ministerkollegen werden mit Argusaugen darüber wachen, dass das Konzept wasserdicht ist. Tatsächlich gibt es noch etliche offene Fragen. So ist etwa unklar, wer Abfindungen erhalten soll, oder was passiert, wenn zu wenig Lehrer die Angebote des vorzeitigen Ausstiegs nutzen. Dennoch gibt es zu Reiches Personalpapier eigentlich keine Alternative. Das Land hat künftig zu wenig Arbeit für zu viele Pädagogen. Und die Zeit drängt. Der große Einbruch der Schülerzahlen an den weiterführenden Schulen beginnt bereits im Herbst 2003. Die Regierung muss sich entscheiden, und zwar schnell. Entweder sparen um jeden Preis - oder Priorität für die Bildung.



Hierarchy of Definiteness

(Gundel et al. 93)

In focus	“it”
Activated	“this”/“that”
Familiar	“this N”, “that N”
Identifiable	“the N”
Referential	indefinite “this N”
Identifiable	“a N”

Salience factors in the RAP algorithm

(Lappin/Leass 94)

Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Non-adverbial emphasis	50
Head noun emphasis	80

Sample run: RAP

Sue found a plastic unicorn in the garden. She handed it to Jill. She liked it very much.

Step	Referent	Referring Expressions	Value
(1)	Sue	{*Sue*}	310
	unicorn	{*a plastic unicorn*}	280
	garden	{*the garden*}	230
(2)	Sue	{*Sue*}	155
	unicorn	{*a plastic unicorn*}	140
	garden	{*the garden*}	115
(3)	Sue	{*Sue, she*}	155+310=465
	unicorn	{*a plastic unicorn, it*}	140+280=420
	garden	{*the garden*}	115
	Jill	{*Jill*}	270
(4)	Sue	{*Sue, she, she*}	232.5
	unicorn	{*a plastic unicorn, it, it*}	210
	garden	{*the garden*}	57.5
	Jill	{*Jill*}	135

Corpus annotation

- Starting with MUC in the 1990s, several text corpora have been annotated with coreference information and used in competitions
  - MUC: message understanding conference
  - DUC: document understanding conference
  - TREC: text retrieval evaluation conference
  - ...

Beware of the task definition

- Competition scenario
  Work with standard data sets (MUC, DUC, TREC, ...): markables already identified – need to decide on coreference only

- Real-world scenario
  Work with authentic raw text, where mentions have to be identified first

- E.g.: Soon et al (01) found that with tagging, chunking, NER, 85% of mentions in the MUC-7 corpus are being detected

Mention-pair Models

- Stage 1: Pairwise classification

© Manfred Stede / NAACL Tutorial 2013

Mention-pair Models: Features
(Soon et al. 01)

- DIST
- SEMCLASS
- NUMBER
- GENDER
- PROPER-NAME
- ALIAS
- ANA-PRONOUN
- DEF-NP
- DEM-NP
- STR-MATCH
- APPOSITIVE
- ANTE-PRONOUN

Mention-pair Models

- Stage 1: Pairwise classification
  - NP1 = NP2
  - NP2 = NP3
  - NP1 =/= NP3

© Manfred Stede / NAACL Tutorial 2013

Mention-pair Models

- Stage 1: Pairwise classification
  - NP1 = NP2
  - NP2 = NP3
  - NP1 =/= NP3
- Stage 2: Clustering
  - Local: consider a small number of pairings
    - Soon et al. 01: „closest first“
  - Global: consider as many pairings as possible

Generating training instances

- Given an annotated corpus, what do we learn from?
  - all pairs of mentions => skewed distribution
  - mirror the human disambiguation task:
    anaphor + antecedent
    anaphor + all „intervening“ mentions

Example

(Soon et al. 01)

Example 3.6 **(Ms. Washington)**$_{73}$'s candidacy is being championed by (several powerful lawmakers)$_{74}$ including ((**her**)$_{76}$ boss)$_{75}$), (Chairman John Dingell)$_{77}$ (D., (Mich.)$_{78}$) of (the House Energy and Commerce Committee)$_{79}$. **(She)**$_{80}$ currently is (a counsel)$_{81}$ to (the committee)$_{82}$.

ante	ana	feature vector	class. decision
(several powerful lawmakers)$_{74}$	(her)$_{76}$	0, 1, −, 2, −, −, +, −, −, −, −, −	no
(Ms. Washington)$_{73}$	(her)$_{76}$	0, 1, +, 1, −, −, +, −, −, −, −, −	yes
(the House Energy and C. Committee)$_{79}$	(She)$_{80}$	1, 0, +, 0, −, −, +, −, −, −, −, −	no
(Mich.)$_{78}$	(She)$_{80}$	2, 0, +, 0, −, −, +, −, −, −, −, −	no
(Chairman J.D.)$_{77}$	(She)$_{80}$	3, 1, +, 0, −, −, +, −, −, −, −, −	no
(her)$_{76}$	(She)$_{80}$	3, 1, +, 1, −, −, +, −, −, −, −, +	yes

Mention-pair models

Many refinements, different classiciation techniques, etc., over the years

Fundamental problem, though:

Jim Miller ... a unicorn ... Laura Smith ... she ... her brother ... Miller ... the beast ... he ...

Incremental entity-mention models

(here: Klenner/Tuggener 2011)

- I: the chronologically-ordered list of mentions
- C: set of coreference sets
- B: buffer for non-anaphoric mentions
- m_i: current mention
- +: concatenate an item to a list

Incremental entity-mention models

Klenner/Tuggener 2011

```
for  i=1      to length(I)
     for      j=1 to length(C)
              r_j := virtual prototype of coreference set C_j
              Cand := Cand ⊕ r_j if compatible(r_j, m_i)
     for      k= length(B) to 1
              b_k:= the k-th licensed buffer element
              Cand := Cand ⊕ b_k if compatible(b_k, m_i)
if   Cand     = {} then B := B ⊕ m_i
if   Cand     ≠ {} then
     ante_i   := most salient element of Cand
     C        := augment(C,ante_i,m_i)
```

One complication: Indirect anaphora / Bridging
(Clark 1977)

- Necessary parts:
  I entered the room. The ceiling was high.
- Probable parts:
  I entered the room. The windows looked out to the bay.
- Inducible parts:
  I entered the room. The chandeliers sparkled brightly.
- Necessary roles:
  I went shopping. The time I started was 3pm.
- Optional roles:
  John was murdered. The knife lay nearby.
- Relations like reason, cause, consequence:
  An earthquake (...). The suffering people are going through (...)

Some more complications...

- Non-nominal antecedents
  Jim met his mother last week. His father didn't like that.
- Reference to sets of objects
  Today I saw two roses and later on three tulips. They all were of the same red.
- Generic readings
  Today I ran into a bunch of squirrels. They are really wonderful animals.
- Expletive "it"
  It was raining the whole day long.
- Cataphora
  Before it ran away, the unicorn looked at me sadly.
- Ellipsis / one-anaphora
  Mike likes black cats, while Paul prefers brown ones.

Tutorial overview

- Part 1: „Large“ discourse units
  - Genre-specific structure
  - Topics
- Part 2: Coreference
- Part 3: „Small“ discourse units
  - Local coherence
  - Coherence-relational text structure
- Exploring interconnections

- John took a train from Paris to Istanbul. He has family there.
- John took a train from Paris to Istanbul. He likes spinach. (Hobbs 79)
- John took a train from Paris to Istanbul. Turkey has become a popular tourist destination.

Coherence relations

- John took a train from Paris to Istanbul. He has family there.
- John took a train from Paris to Istanbul. He likes spinach.
- John took a train from Paris to Istanbul. Turkey has become a popular tourist destination.
- John took a train from Paris to Istanbul, because he has family there.

Coherence relations

- John took a train from Paris to Istanbul, but he never arrived.
- Although John took a fast train from Paris to Istanbul, he arrived late.
- ...

Coherence relations: levels of description

- If you‘re thirsty, make sure that you find some water.
  SEMANTICS

- If you‘re thirsty, there‘s a beer in the fridge.
  PRAGMATICS

Definition

- Coherence relation

  A relationship between adjacent units of text, holding on a semantic (proposition) or pragmatic (speech act) level of analysis. Common groupings of relations are
  - causality
  - similarity/contrast
  - contiguity

Issues

- Connectives
- Minimal units of the analysis
- Inventory and definitions of coherence relations
- Corpus: PDTB
- Automatic local coherence analysis

- Coherence relations and discourse structure
- Corpora: RST-DT, Discourse Graph Bank
- RST-parsing

Connectives

if you use a character reference such as < to insert the < character, the formatter will output <.

if you use a character reference such as < to insert the < character, the formatter will output <.

Coherence relation:

Purpose (you use &#60 , you insert <)

if you use a character reference such as < to insert the < character, the formatter will output <.

Coherence relation:

Condition (you use &#60 to insert < ,
formatter will output <)

Because well-formed XML does not permit raw less-than signs and ampersands, if you use a character reference such as < or the entity reference < to insert the < character, the formatter will output < or perhaps <.

nonetheless
notwithstanding that
all the same
despite the fact that
though
although
rather
however
but
or else
on the other hand
on the contrary
only
even so
in spite of that
in contrast
else
despite this
yet
instead
failing that
on the other side
unless
then again
whereas
admittedly
nevertheless
apart from that
whilst
by contrast
otherwise
alternatively
while
anyway
meanwhile
still
even though

(from Knott 96)

Definition

- Connective

  A closed-class, non-inflectable word or word group that semantically denotes a two-place relation, where the entities being related can in text be expressed as clauses.

  Syntactically, a connective can be
  - a subordinating conjunction,
  - a coordinating conjunction,
  - an adverbial,
  - (arguably) a preposition.

Connectives: Ambiguity

- Connective or no connective
  - *Since you like ice cream so much, I'll buy one for you.*
  - *Since 1988 I have never had any ice cream.*

- Different coherence relations
  - *Since you like ice cream so much, I'll buy one for you.*
  - *I haven't had any ice cream since you arrived in New York.*

Issues

- Connectives
- Minimal units of the analysis
- Inventory and definitions of coherence relations
- Corpus: PDTB
- Automatic local coherence analysis

- Coherence relations and discourse structure
- Corpora: RST-DT, Discourse Graph Bank
- RST parsing

Minimal units

- Sentences!
  - But:
    - incomplete material, ellipsis
      Bring me a hammer please. A big one.
    - complex sentences
      Bring me a hammer when you pass by the workshop.
- Clauses!
  - But:
    - deal with nominalizations
      John attended the lecture despite his illness.
    - deal with embedding
      John, although he was ill, attended the lecture.
    - deal with non-/restrictive relative clauses
      The red car that's parking in front of you belongs to me.
      The red car, which was brought here by my Dad, belongs to me.

Minimal units

- Non-structural definitions, e.g. Polanyi et al. (04):
  1) „the syntactic constructions that encode a minimum unit of meaning and/ or discourse function interpretable relative to a set of contexts."

Table 4.1: Examples of minimal discourse units and their types [Polanyi et al., 2004].

Segment types	Realizations	Examples
Eventualities	clauses	[I heard the dog][that was barking.]
	predication	[California elected Schwarzenegger] [governor.]
	infinitival modif.	[They left] [to get the tickets.]
Interpolations	parentheticals	[The show [(and what a show it was)] lasted 4 hours.]
Fragments	section headings	[4. Discussion]
	list items	[e.g., [hydrogen,] [helium]]
Conj. operators	conjunction	[We arrived] [and] [got seats.]
Discourse operators	"scene setting" modifier	[On Tuesday,] [we will see the sites.]

2) units that can be „independently continued"
   (removes many small units from the candidate set)

Definition

- Elementary discourse unit (EDU)

  A span of text, usually a clause, but in general ranging from minimally a (nominalization) NP to maximally a sentence. It denotes a single event or type of event, serving as a complete, distinct unit of information that the surrounding discourse may connect to. An EDU may be structurally embedded in another.

Automatic segmentation

- Example: SLSeg (Tofiloski et al. 2009)
- Human agreement: kappa 0.85
- Based on syntactic rules, e.g.
  - distinguish sentential complements that should/not be EDUs
  - copy heads of non-restrictive relative clauses

>The aftermath of the 2008 cyclone in Burma not only betrayed the callous indifference of the ruling junta
>but demonstrated the vibrancy of civil society there.
>Haiti's earthquake shows that, whatever the communal spirit of its people at the moment of crisis,
>the government was not functioning, unable even to bury the dead, much less rescue the living.
>This vacuum, which had been temporarily filled by the U.N.,
>This vacuum now poses the threat of chaos.

www.sfu.ca/˜mtaboada/research/SLSeg.html

Issues

- Connectives
- Minimal units of the analysis
- Inventory and definitions of coherence relations
- Corpus: PDTB
- Automatic local coherence analysis

- Coherence relations and discourse structure
- Corpora: RST-DT, Discourse Graph Bank
- RST parsing

Inventory and definitions of relations

- Many proposals in the (discourse, semantics, pragmatics) literature
- Three influential ones:
  - Segmented Discourse Representation Theory (Asher/Lascarides 03)
  - Rhetorical Structure Theory (Mann/Thompson 88)
  - Connective senses in Penn Discourse Treebank (Prasad et al. 08)

Relation definitions

- Evidence in RST:
  - *Constraint on Nucleus*:
    Reader might not believe N to a degree satisfactory to Writer
  - *Constraint on Satellite*:
    Reader believes S or will find it credible
  - *Constraints on Nucleus+Satellite*:
    Reader's comprehending S increases her/his belief of N
  - *Intention of Writer*:
    Reader's belief of N is increased

TEMPORAL
– Asynchronous
– Synchronous
—— precedence
—— succession

CONTINGENCY
– Cause
—— reason
—— result
– Pragmatic Cause
—— justification
– Condition
—— hypothetical
—— general
—— unreal present
—— unreal past
—— factual present
—— factual past
– Pragmatic Condition
—— relevance
—— implicit assertion

COMPARISON
– Contrast
—— juxtaposition
—— opposition
– Pragmatic Contrast
– Concession
—— expectation
—— contra-expectation
– Pragmatic Concession

EXPANSION
– Conjunction
– Instantiation
– Restatement
—— specification
—— equivalence
—— generalization
– Alternative
—— conjunctive
—— disjunctive
—— chosen alternative
– Exception
– List

Figure 4.7: Hierarchy of connective senses in the PDTB [Prasad et al., 2008].

Corpus: PDTB (Prasad et al. 08)

- Wall Street Journal portion of Penn Treebank
- 100 connectives (types)
- 18.000 instances
- Implicit connectives: Within paragraphs, check whether a connective *could* be added between segments

Example 4.26 [Drug makers shouldn't be able to duck liability]$_{Arg1}$ [because]$_{Conn}$ [people couldn't identify precisely which identical drug was used.]$_{Arg2}$

Example 4.27 [France's second-largest government-owned insurance company, Assurances Generales de France, has been building its own Navigation Mixte stake]$_{Arg1}$ currently thought to be between 8% and 10%. Analysts said [they don't think it is contemplating a takeover]$_{Arg2}$, [however]$_{Conn}$, and its officials couldn't be reached.

- Agreement:
  - identify both arguments of explicit connective: 90.2%
  - identify both arguments of implicit connective: 85.1%
  - identify the relation at top-level (4 groups): 94%

Automatic local coherence analysis

- E.g., as an extension of opinion mining:

  (...) *We didn't like the village very much and won't come back because there are only few things to see, and moreover the place is quite dirty. (...)*

Automatic local coherence analysis

- 1 - Disambiguate potential connectives
  - Marcu 00: 1200 of 2100 potential cue phrases *are* cue phrases
  - Pitler/Nenkova 09
    - Of 100 candidates in PDTB, only 11 appear as connectives more than 90% of the time
    - MaxEnt classifier using syntactic features (from manual annotation): f-measure 92.3%
      - POS tag, phrase label of candidate
      - categories of parent and left sibling
      - right sibling: catgory, presence of VP or trace in subtree

Automatic local coherence analysis

- 2 – Sense disambiguation (PDTB-style) for connectives

- Miltsakaki et al. 05
  - *since* (temporal/causal): 89.5% acc.
  - *while* (temporal/opposition/concessive): 71.9%
  - *when* (temporal/conditional): 82.6%
  - Features: tense form of aux. *have* and *be*, tense form of head, presence of modals and temp. expr.
- Pitler/Nenkova 09
  - due to skewed sense distribution, majority class leads to 93.7% acc. already; with syntactic feaures: 94.2%

Automatic local coherence analysis

- 3 – Scope identification

- **Preposition**: Arg2 is the NP following the preposition, Arg1 is the governing clause.
  Despite [her bad mood,]$_{Arg2}$ [Susan decided to go to the party.]$_{Arg1}$
- **Subordinating conjunction:** Arg2 is the clause following the conjunction, Arg1 is the matrix clause.
  Because [Susan was in a bad mood,]$_{Arg2}$ [she did not go to the party.]$_{Arg1}$
- **Coordinating conjunction**: Arg2 is the sentence following the conjunction, Arg1 is some text segment preceding it.
  [Susan was in a bad mood,]$_{Arg1}$ but [she decided to go to the party.]$_{Arg2}$
- **Adverbial:** Arg2 is the clause containing the adverbial; Arg1 is some text segment preceding it.
  [Susan was in a bad mood.]$_{Arg1}$ [She nevertheless decided to go to the party.]$_{Arg2}$

Automatic local coherence analysis

- 3 – Scope identification
- Elwell/Baldridge 08
  - Arg2: acc. 92-94%
  - Arg1: acc. 78-82%
- [Drug makers shouldn't be able [to duck liability]]$_{Arg1?}$ because [people couldn't identify precisely which identical drug was used.]$_{Arg2}$
- Prasad et al. 08: Arg1 distribution in PDTB
  - 65% in same sentence
  - 30% in immediately-preceding sentence
  - 9% in non-adjacent sentence

Automatic local coherence analysis

- 4 – Finding implicit relations

- Proportion of signalled relations
  - roughly 40% (according to several studies)
  - differs considerably between relations! E.g., CONCESSION versus BACKGROUND

Automatic local coherence analysis

- 4 – Finding implicit relations

- Sporleder/Lascarides 05
  - 5-way classification: CONTRAST, EXPLANATION, RESULT, SUMMARY, CONTINUATION
  - Approach: Remove connective, then classify
  - Features: position, length of arguments, words, parts of speech, WordNet classes, tens/aspect and various syntactic f.s
  - Acc.: 57.6%

  - (Assumption made: contexts of explicit and implicit instances are similar to each other – but are they? See Sporleder/Lascarides 08)

A final version of the train example...

- John took a train from Paris to Istanbul, departing from Montparnasse at noon. He has family in Istanbul.

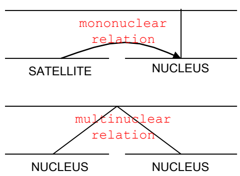
Issues

- Connectives
- Minimal units of the analysis
- Inventory and definitions of coherence relations
- Corpus: PDTB
- Automatic local coherence analysis

- Coherence relations and discourse structure
- Corpora: RST-DT, Discourse Graph Bank
- RST parsing

Rhetorical Structure Theory (RST; Mann/Thompson 88)

- Coherence relations between discourse segments
  - asymmetric ("mononuclear")
    - one nucleus, one satellite
  - symmetric ("multinuclear")
    - multiple nuclei

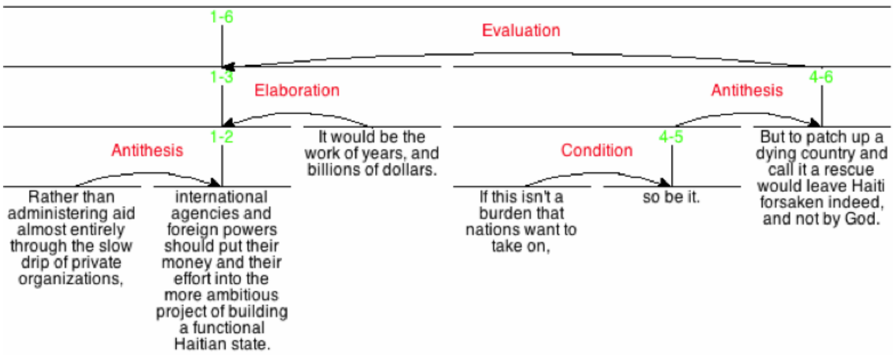


- Resulting structure is a complete tree
  - No gaps
  - No cross-dependencies

- 25 relations
  Cause, Contrast, Elaboration, ...

© Manfred Stede / NAACL Tutorial 2013

Rhetorical Structure Theory: Example



RST Discourse Treebank

(Carlson et al. 03)

- 385 Wall Street Journal articles (22k EDUs)
- 53 mononuclear, 25 multinuclear relations
- Relations grouped into 16 categories
- Detailed annotation guidelines; kappa for the 16 categories up to .82

- Attribution
- Background
- Cause
- Comparison
- Condition
- Contrast
- Elaboration
- Enablement
- Evaluation
- Explanation
- Joint
- Manner-Means
- Topic-Comment
- Summary
- Temporal
- TopicChange

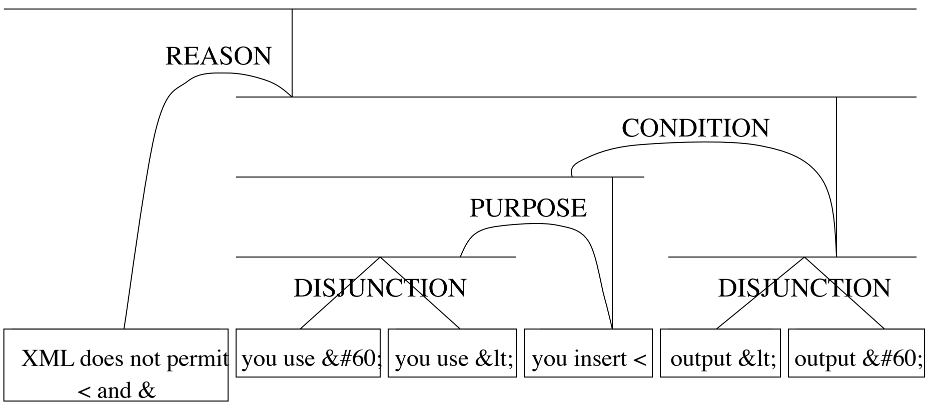
Structure

Because well-formed XML does not permit raw less-than signs and ampersands, if you use a character reference such as < or the entity reference < to insert the < character, the formatter will output < or perhaps <.

Because A, if B or C to D, E or F.

(Because A, (if ((B or C) to D), (E or F))).

„Rhetorical structure“



Greedy RST parsing (Hernault et al. 10)

input: $L \leftarrow < e_1, e_2, ..., e_n >$ (list of EDUs)
for all (l_i, l_{i+1}) in L **do**
$Scores[i] \leftarrow \text{STRUCT}(l_i, l_{i+1})$
end for
while $|L| > 1$ **do**
$i \leftarrow max(Scores)$
$NewLabel \leftarrow \text{LABEL}(l_i, l_{i+1})$
$NewSubTree \leftarrow \text{CreateTree}(l_i, l_{i+1}, NewLabel)$
$Scores[i-1] \leftarrow \text{STRUCT}(l_{i-1}, NewSubTree)$
$Scores[i+2] \leftarrow \text{STRUCT}(NewSubTree, l_{i+2})$
$delete(Scores[i])$
$delete(Scores[i+1])$
$L \leftarrow [l_0, ..., l_{i-1}, NewSubTree, l_{i+2}, ...]$
end while
return l_0

Greedy RST parsing (Hernault et al. 10)

- SVMs trained on RST-DT
- STRUCT: binary classifier scoring whether *any* relation holds between a pair of segments
  - features: sentence and para boundaries, segment size, position in text
- LABEL: multi-class classifier assigning relation and nuclearity assignment
  - features similar to those discussed earlier
  - plus relations that have already been assigned

Trees or Graphs?

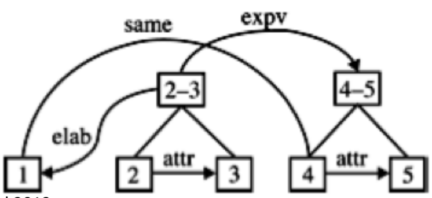
- Webber et al. 99: structural versus anahoric connectives

  (a) *On the one hand, John loves Barolo.* (b) *So he ordered three cases of the `97.* (c) *On the other hand, because he's broke,* (d) *he then had to cancel the order.*

- See also: Egg/Redeker (10)

Discourse Graph Bank (Wolf/Gibson 05)

- 135 texts: AP newswire, Wall Street Journal
- Distinguish directed from undirected relations
- No nuclearity
- Example: (1) *Mr. Baker's assistant for inter-American affairs, 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."*



EDUs and relations: where are we?

- RST parsing: technically interesting (RST-DT), but open issues w.r.t. theoretical status
  - what relations?
  - intentional or informational analysis?
  - what is the right unit for RST tree? paragraph?
- PDTB: much less commitment on overall discourse structure – very useful for local coherence anaylsis

Tutorial overview

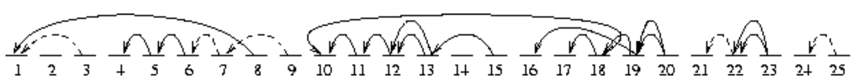
- Part 1: „Large“ discourse units
  - Genre-specific structure
  - Topics
- Part 2: Coreference
- Part 3: „Small“ discourse units
  - Local coherence
  - Coherence-relational text structure
- Exploring interconnections

© Manfred Stede / NAACL Tutorial 2013

Mandatory vaccination against children's diseases?

[1] Today, children don't know anymore what pox are. [2] What a joy. [3] When pox vaccination was introduced in 1854, [4] quite a few people believed [5] that their head would turn into a cow's head [6] if they got themselves vaccinated. [7] For the vaccine was made from cattle's skin at the times. [8] Nowadays this dreadful disease is exterminated. [9] Thanks to a determined, world-wide vaccination campaign. [10] But there still are other diseases: Measles, polio, diphteria, mumps, rubella, hepatitis B, tuberculosis, pertussis. [11] Millions of children die of these, especially in less developed countries. [12] In Germany, many parents apparently don't take these diseases seriously. [13] Because they don't know them anymore! [14] For it has been achieved with vaccines [15] that these infections hit only rarely today. [16] But those who have experienced [17] how terribly children suffer [18] when they come down with ‚just' measles or pertussis, [19] should spare them the agony. [20] As well as the long-term consequences. [21] Only those who have their children vaccinated will contribute to vaccines' becoming superfluous some day. [22] Instead, people rant about side effects [23] that occur very rarely and are known merely from books. [24] Then there is the great argument: This is my child, the governement must not prick her. [25] No vaccine can help against such parents.

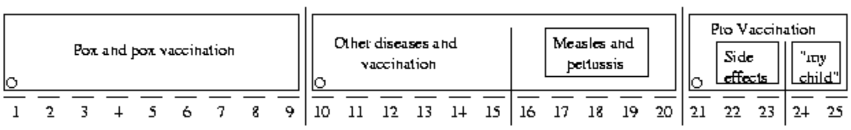
Referential Structure



Mandatory vaccination against children's diseases?

[1] Today, children don't know anymore what pox are. [2] What a joy. [3] When pox vaccination was introduced in 1854, [4] quite a few people believed [5] that their head would turn into a cow's head [6] if they got themselves vaccinated. [7] For the vaccine was made from cattle's skin at the times. [8] Nowadays this dreadful disease is exterminated. [9] Thanks to a determined, world-wide vaccination campaign. [10] But there still are other diseases: Measles, polio, diphteria, mumps, rubella, hepatitis B, tuberculosis, pertussis. [11] Millions of children die of these, especially in less developed countries. [12] In Germany, many parents apparently don't take these diseases seriously. [13] Because they don't know them anymore! [14] For it has been achieved with vaccines [15] that these infections hit only rarely today. [16] But those who have experienced [17] how terribly children suffer [18] when they come down with ‚just' measles or pertussis, [19] should spare them the agony. [20] As well as the long-term consequences. [21] Only those who have their children vaccinated will contribute to vaccines' becoming superfluous some day. [22] Instead, people rant about side effects [23] that occur very rarely and are known merely from books. [24] Then there is the great argument: This is my child, the governement must not prick her. [25] No vaccine can help against such parents.

Topic Structure



Mandatory vaccination against children's diseases?

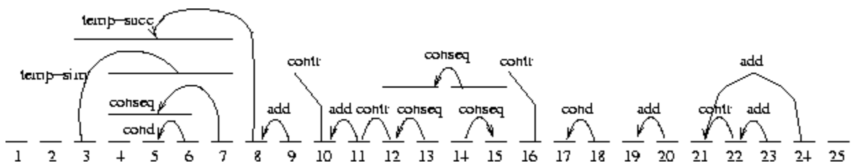
[1] Today, children don't know anymore what pox are. [2] What a joy. [3] When pox vaccination was introduced in 1854, [4] quite a few people believed [5] that their head would turn into a cow's head [6] if they got themselves vaccinated. [7] For the vaccine was made from cattle's skin at the times. [8] Nowadays this dreadful disease is exterminated. [9] Thanks to a determined, world-wide vaccination campaign. [10] But there still are other diseases: Measles, polio, diphteria, mumps, rubella, hepatitis B, tuberculosis, pertussis. [11] Millions of children die of these, especially in less developed countries. [12] In Germany, many parents apparently don't take these diseases seriously. [13] Because they don't know them anymore! [14] For it has been achieved with vaccines [15] that these infections hit only rarely today. [16] But those who have experienced [17] how terribly children suffer [18] when they come down with ‚just' measles or pertussis, [19] should spare them the agony. [20] As well as the long-term consequences. [21] Only those who have their children vaccinated will contribute to vaccines' becoming superfluous some day. [22] Instead, people rant about side effects [23] that occur very rarely and are known merely from books. [24] Then there is the great argument: This is my child, the governement must not prick her. [25] No vaccine can help against such parents.

Conjunctive Relations

- temporal
  - simultaneous, succession
- consequential
  - manner, consequence, condition, purpose, concession
- comparative
  - similarity, contrast, reformulation
- additive
  - addition, alternation

(Martin 1992); see also: PDTB

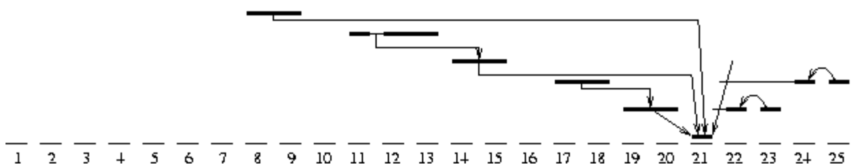
Conjunctive Relations



Mandatory vaccination against children‘s diseases?

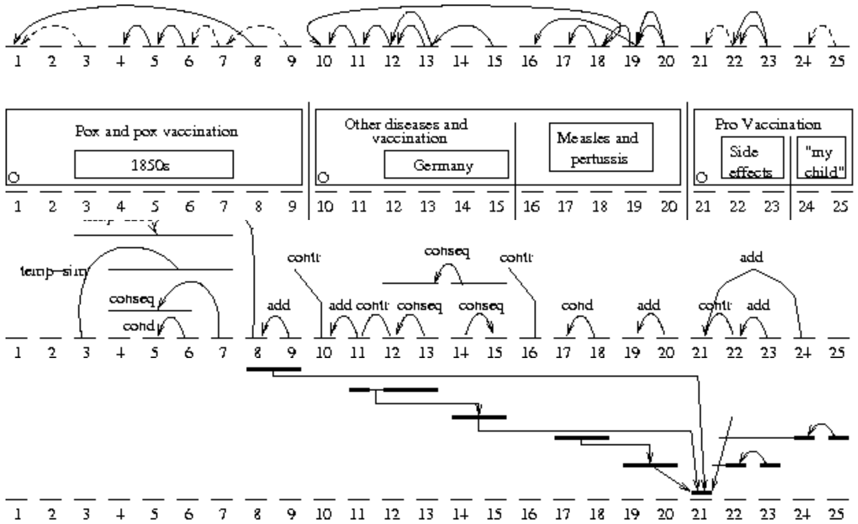
[1] Today, children don‘t know anymore what pox are. [2] What a joy. [3] When pox vaccination was introduced in 1854, [4] quite a few people believed [5] that their head would turn into a cow‘s head [6] if they got themselves vaccinated. [7] For the vaccine was made from cattle‘s skin at the times. [8] Nowadays this dreadful disease is exterminated. [9] Thanks to a determined, world-wide vaccination campaign. [10] But there still are other diseases: Measles, polio, diphteria, mumps, rubella, hepatitis B, tuberculosis, pertussis. [11] Millions of children die of these, especially in less developed countries. [12] In Germany, many parents apparently don‘t take these diseases seriously. [13] Because they don‘t know them anymore! [14] For it has been achieved with vaccines [15] that these infections hit only rarely today. [16] But those who have experienced [17] how terribly children suffer [18] when they come down with ‚just‘ measles or pertussis, [19] should spare them the agony. [20] As well as the long-term consequences. [21] Only those who have their children vaccinated will contribute to vaccines‘ becoming superfluous some day. [22] Instead, people rant about side effects [23] that occur very rarely and are known merely from books. [24] Then there is the great argument: This is my child, the governement must not prick her. [25] No vaccine can help against such parents.

Argument structure



(inspired by Freeman 1993)

Text understanding: Relating levels of analysis

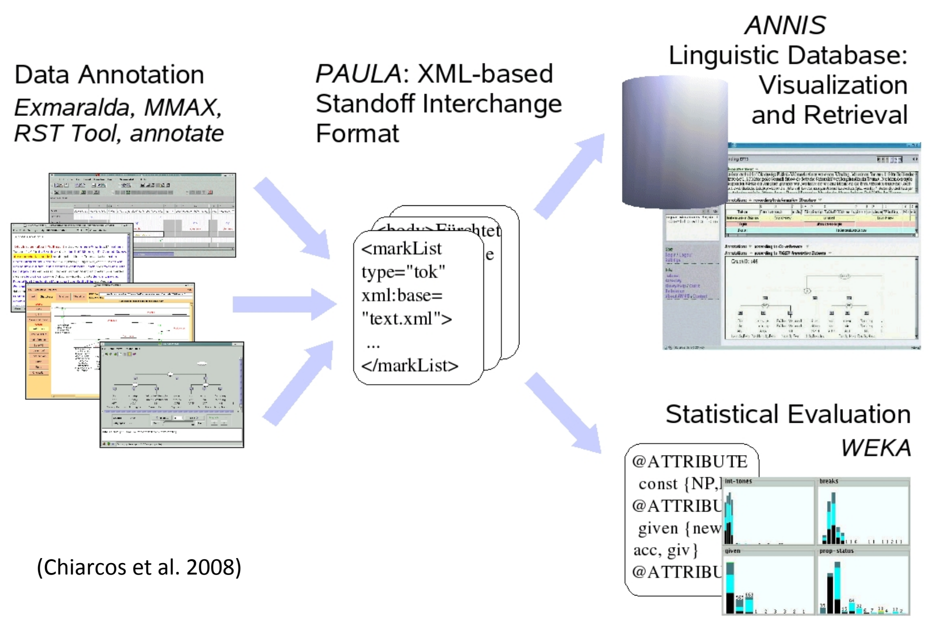


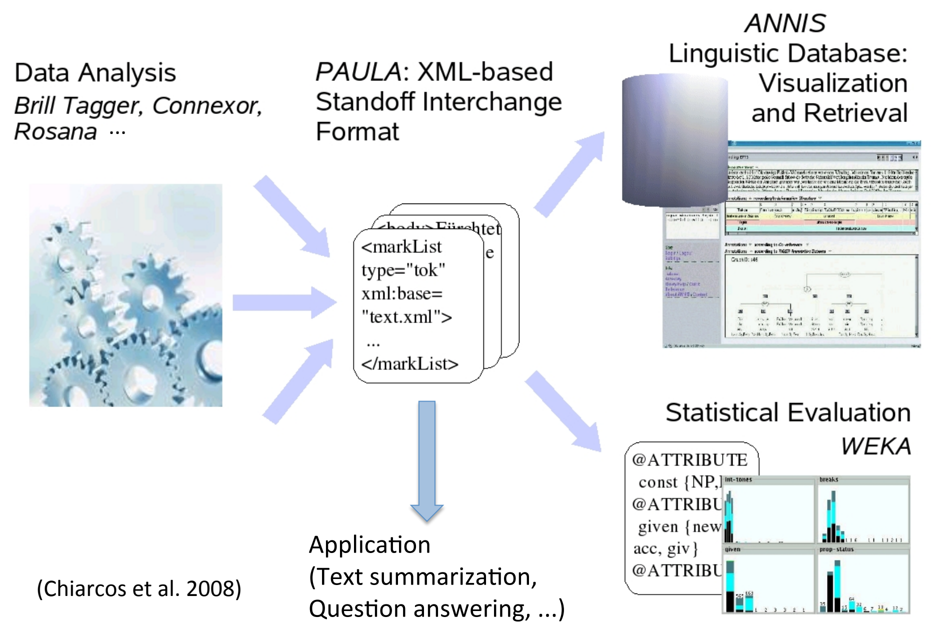
Potsdam Commentary Corpus (PCC)
(Stede 2004)

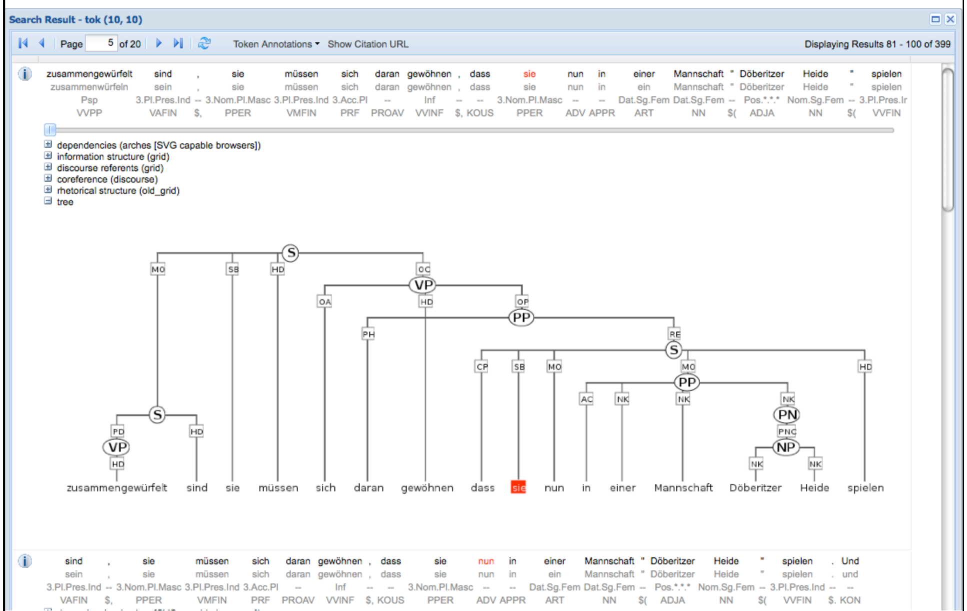
200 short editorials from two German newspapers

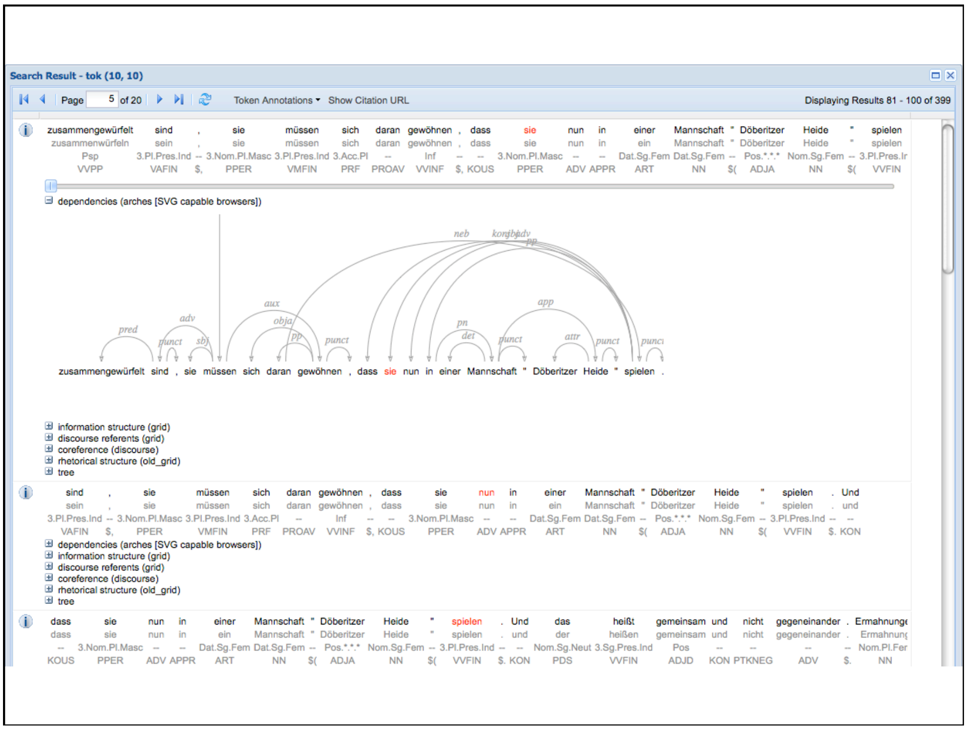
- Syntax
- Coreference
- RST

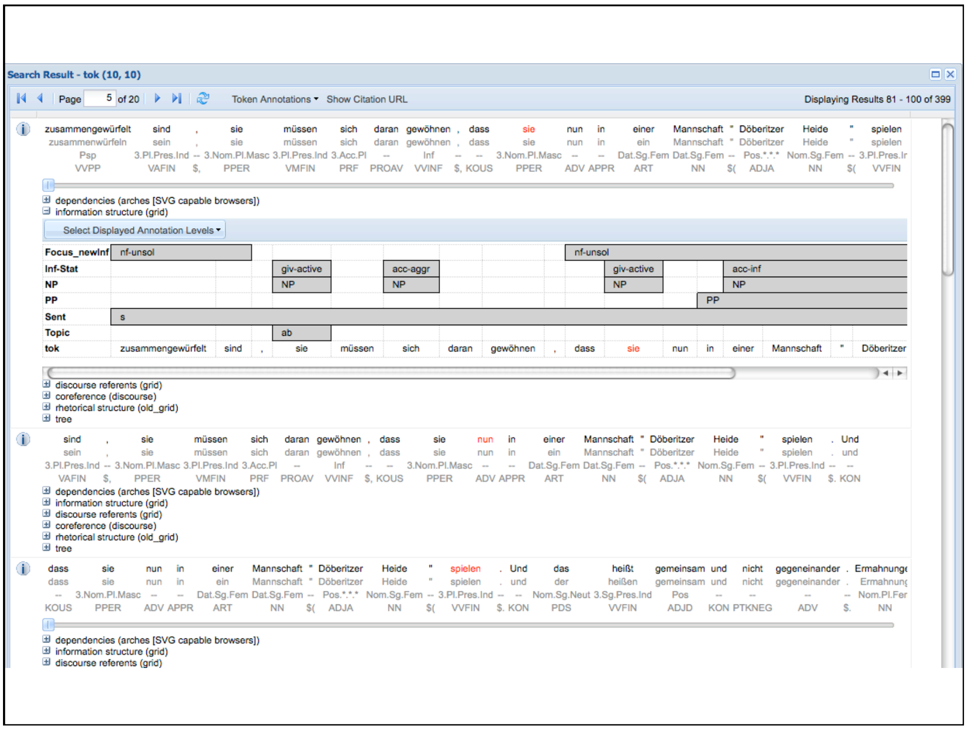
- Content zones
- Topics
- Information structure
- Conjunctive relations
- Illocutionary status
- Argumentation structure

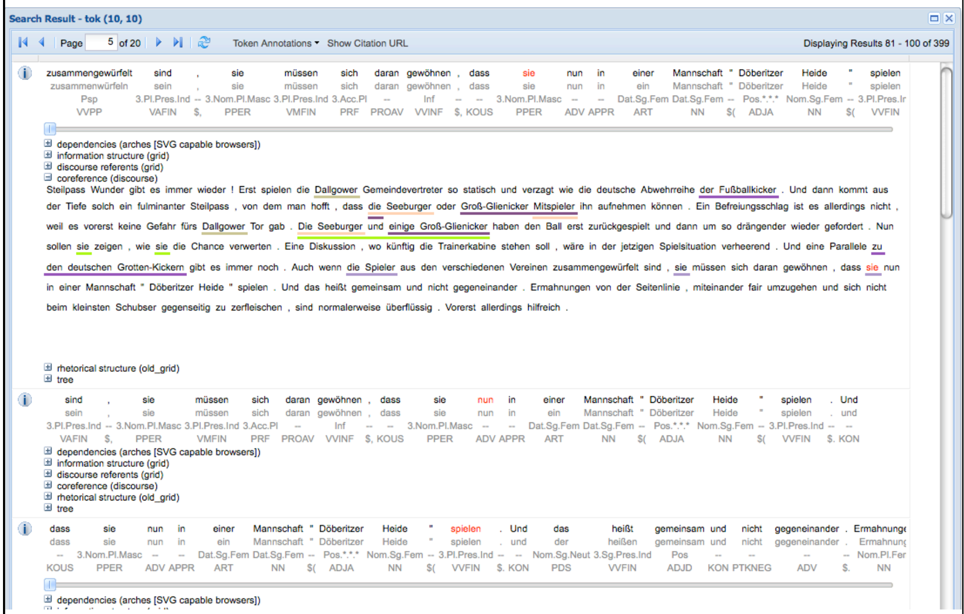



Data Analysis
Brill Tagger, Connexor, Rosana …
PAULA: XML-based Standoff Interchange Format
<markList type="tok" xml:base= "text.xml">
…
</markList>
ANNIS
Linguistic Database: Visualization and Retrieval
Statistical Evaluation
WEKA
@ATTRIBUTE const {NP,
@ATTRIBU
given {new
acc, giv}
@ATTRIBU
Application
(Text summarization, Question answering, ...)
(Chiarcos et al. 2008)


Search Result - tok (10, 10)
Page 5 of 20
Token Annotations
Show Citation URL
Displaying Results 81 - 100 of 399
dependencies (arches [SVG capable browsers])
information structure (grid)
discourse referents (grid)
coreference (discourse)
rhetorical structure (old_grid)
tree
zusammengewürfelt sind sie müssen sich daran gewöhnen dass sie nun in einer Mannschaft Döberitzer Heide spielen


Search Result - tok (10, 10)
Page 5 of 20
Token Annotations
Show Citation URL
Displaying Results 81 - 100 of 399
dependencies (arches [SVG capable browsers])
zusammengewürfelt sind , sie müssen sich daran gewöhnen , dass sie nun in einer Mannschaft " Döberitzer Heide " spielen .
information structure (grid)
discourse referents (grid)
coreference (discourse)
rhetorical structure (old_grid)
tree


Search Result - tok (10, 10)
Page 5 of 20
Token Annotations
Show Citation URL
Displaying Results 81 - 100 of 399
dependencies (arches [SVG capable browsers])
information structure (grid)
Select Displayed Annotation Levels
Focus_newInf
Inf-Stat
NP
PP
Sent
Topic
tok
discourse referents (grid)
coreference (discourse)
rhetorical structure (old_grid)
tree



ANNIS2 (Zeldes et al. 09)

- ExtJS (JavaScript) web interface, PostGreSQL backend
- Types of annotations that can be merged:
  - spans with labels
  - DAGs with labelled edges
  - pointing relations between terminals and non-terminals
- Query language
  - exact match, regexp match on primary data and annotations
  - relations between elements
    - overlapping/contained/adjacent spans
    - hierarchical dominance (direct/indirect, common ancestors, left-/rightmost child,edges with specific labels, etc.)
- Search across different layers (using namespaces)
- Imported „standard" corpora: TIGER, TüBa/D-Z, Ontonotes, ...
- http://www.sfb632.uni-potsdam.de/~d1/annis/

Example: multi-level querying

Pronominal anaphors with subject antecedent in an RST-satellite, please. (ProCon10 corpus)

```
mmax:node &                                    #1: anaphor
tiger:pos=/P.*/ & #1 _=_ #2 &                  #2: anaphor is pronoun
mmax:node & #1 ->anaph_antec #3 &              #3: antecedent (coref)
tiger:node & #4 _=_ #3 &                       #4: antecedent (syn)
tiger:cat=„S" & #5 >[func=„SB"]#4 &            #5: antec. sentence
                                                   with #4 as subject
rst:cat=„segment" & #6 _i_ #4 &                #6: antec. edu incl. #4
rst:cat & #7 >[func=/.*SAT/] #6                #7: parent of antec. edu
                                                   is satellite
```


Example: multi-level querying

Pronominal anaphors with subject antecedent in an RST-satellite, please. (ProCon10 corpus)

```
mmax:node &                                    #1: anaphor
tiger:pos=/P.*/ & #1 _=_ #2 &                  #2: anaphor is pronoun
mmax:node & #1 ->anaph_antec #3 &              #3: antecedent (coref)
tiger:node & #4 _=_ #3 &                       #4: antecedent (syn)
tiger:cat=„S" & #5 >[func=„SB"]#4 &            #5: antec. sentence
                                                   with #4 as subject
rst:cat=„segment" & #6 _i_ #4 &                #6: antec. edu incl. #4
rst:cat & #7 >[func=/.*SAT/] #6                #7: parent of antec. edu
                                                   is satellite
```


Expectation: pronouns prefer subject antecedents and nucleus antecedents

Das Bild kann nicht angezeigt werden. Dieser Computer verfügt möglicherweise über zu wenig Arbeitsspeicher, um das Bild zu öffnen, oder das Bild ist beschädigt. Starten Sie den Computer neu, und öffnen Sie dann erneut die Datei. Wenn weiterhin das rote x angezeigt wird, müssen Sie das Bild möglicherweise löschen und dann erneut einfügen.

(Chiarcos, subm.)

Accessibility in discourse: Distance-based approaches
(Krasavina/Chiarcos 05)

- Referential Distance

  the number of clauses between anaphor and antecedent has an effect on the form of the anaphor (cf. Givón 1983)

- But

  It is not simple distance that triggers the use of one anaphoric device over the other. Rather, it is the rhetorical organisation of that distance that determines whether a pronoun or a full NP is appropriate. (Fox 1987)

⇒ „Rhetorical Distance"

Thank you for your attention!

References

- Nicholas Asher and Alex Lascarides. Logics of Conversation. Cambridge University Press, Cambridge, 2003
- Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. Machine Learning, 34 (1-3):177–210, 1999
- Douglas Biber. A typology of English texts. Linguistics, 27:3–43, 1989.
- H. Bieler, S. Dipper, M. Stede. *Identifying formal and functional zones in film reviews*. In: Proc. SIGdial WS, Antwerpen, 2007
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, Current Directions in Discourse and Dialogue. Kluwer, Dordrecht, 2003
- C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, M. Stede. *A flexible framework for integrating annotations from different tools and tagsets*. In: Traitement Automatique des Langues, 49:217-246, 2008
- Freddy Choi. Advances in domain independent linear text segmentation. In Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 26–33, Seattle/WA, 2000
- Herbert Clark. Inferences in comprehension. In D. Laberge and S. J. Samuels, editors, Basic Processes in Reading: Perception and Comprehension, pages 243–263. Lawrence Erlbaum, Hillsdale/NJ, 1977
- Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In Proc. of the 22nd national conference on Artificial intelligence (AAAI), pages 1334–1339, 2007
- Markus Egg and Gisela Redeker. How complex is discourse structure? In Proc. of the 7th Conference on International Language Resources and Evaluation (LREC), Malta, 2010
- Jacob Eisenstein. Hierarchical text segmentation from multi-scale lexical cohesion. In Proceedings of ACL/HLT 2009, pages 353–361, Boulder, Colorado, 2009.
- Robert Elwell and Jason Baldridge. Discourse connective argument identification with connective specific rankers. In Proc. of the IEEE Conference on Semantic Computing (ICSC), Santa Clara/CA, 2008
- Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse seg- mentation of multi-party conversation. In Proc. of ACL 2003, pages 562–569, Sapporo, Japan, 2003
- Gundel, Jeanette K., Nancy Hedberg, Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. Language 69:274-307

- Marti A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23 (1):33–64, 1997
- Hugo Hernault, Hemut Prendinger, David duVerle, and Mitsuru Ishizuka. Hilda: A discourse parser using support vector machine classification. Dialogue and Discourse, 1(3):1–33, 2010.
- Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, WordNet: An electronic lexical database. MIT Press, 1998
- Jerry Hobbs. Coherence and coreference. Cognitive Science, 3:67–90, 1979.
- Manfred Klenner and Don Tuggener. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP), pages 178–185. Hissar, Bulgaria, 2011
- Alistair Knott and Robert Dale. Using linguistic phenomena to motivate a set of coherence relations. Discourse Processes, 18(1):35–62, 1994
- Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. Compu- tational Linguistics, 20(4):535–561, 1994
- William Mann and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. TEXT, 8:243–281, 1988
- Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. Computational Linguistics, 26(3):395–448, 2000
- James R. Martin. English text: system and structure. John Benjamins, Philadelphia/Amsterdam, 1992.
- Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, 2005
- Hemant Misra, Francois Yvon, Joemon Jose, and Olivier Cappe. Text segmentation via topic mod- eling: an analytical study. In Proc. of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, 2009
- Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21–48, 1991
- Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL 2004, Barcelona

- Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In Proc. of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore, 2009
- Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. A rule based approach to discourse parsing. In Proc. of the SIGDIAL '04 Workshop, Cambridge/MA, 2004
- R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The Penn Discourse Treebank 2.0. In Proc. of the 6th LREC, Marrakech, Morocco, 2008
- Jeffrey C. Reynar. An automatic method of finding topic boundaries. In Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics, pages 331–333, Las Cruces/NM, June 1994
- H. Gregory Silber and Kathleen F. McCoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics, 28(4):488–496, 2002
- Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4):521–544, 2001
- Caroline Sporleder and Alex Lascarides. Exploiting linguistic cues to classify rhetorical relations. In Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP), 2005.
- Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. Natural Language Engineering, 14(3):369–416, 2008
- Manfred Stede. The Potsdam Commentary Corpus. In Proc. of the ACL Workshop on Discourse Annotation, pages 96–102, Barcelona, 2004
- Nicola Stokes, Joe Carthy, and Alan F. Smeaton. Select: a lexical cohesion based news story seg- mentation system. AI Communication, 17:3–12, 2004
- Maite Taboada, Julian Brooke, and Manfred Stede. Genre-based paragraph classification for sen- timent analysis. In Proc. of the SIGDIAL 2009 Conference, page 62–70, London, UK, September 2009
- M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede. *Lexicon-based methods for sentiment analysis* In: Computational Linguistics, 37(2):267-307, 2011
- S. Teufel, A. Siddharthan, C. Batchelor. 2009. Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In: Proceedings of EMNLP-09, Singapore
- Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In Proc. of ACL/IJCNLP (short papers), Suntec, Singapore, 2009
- Bonnie Webber, Alistair Knott, and Aravind Joshi. Multiple discourse connectives in a lexicalized grammar for discourse. In 3rd International Workshop on Computational Semantics (IWCS), NL- Tilburg, 1999
- Florian Wolf and Edward Gibson. Representing discourse coherence: a corpus-based study. Com- putational Linguistics, 31(2):249–287, 2005
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian (2009), "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In: *Proceedings of Corpus Linguistics 2009*, July 20-23, Liverpool, UK