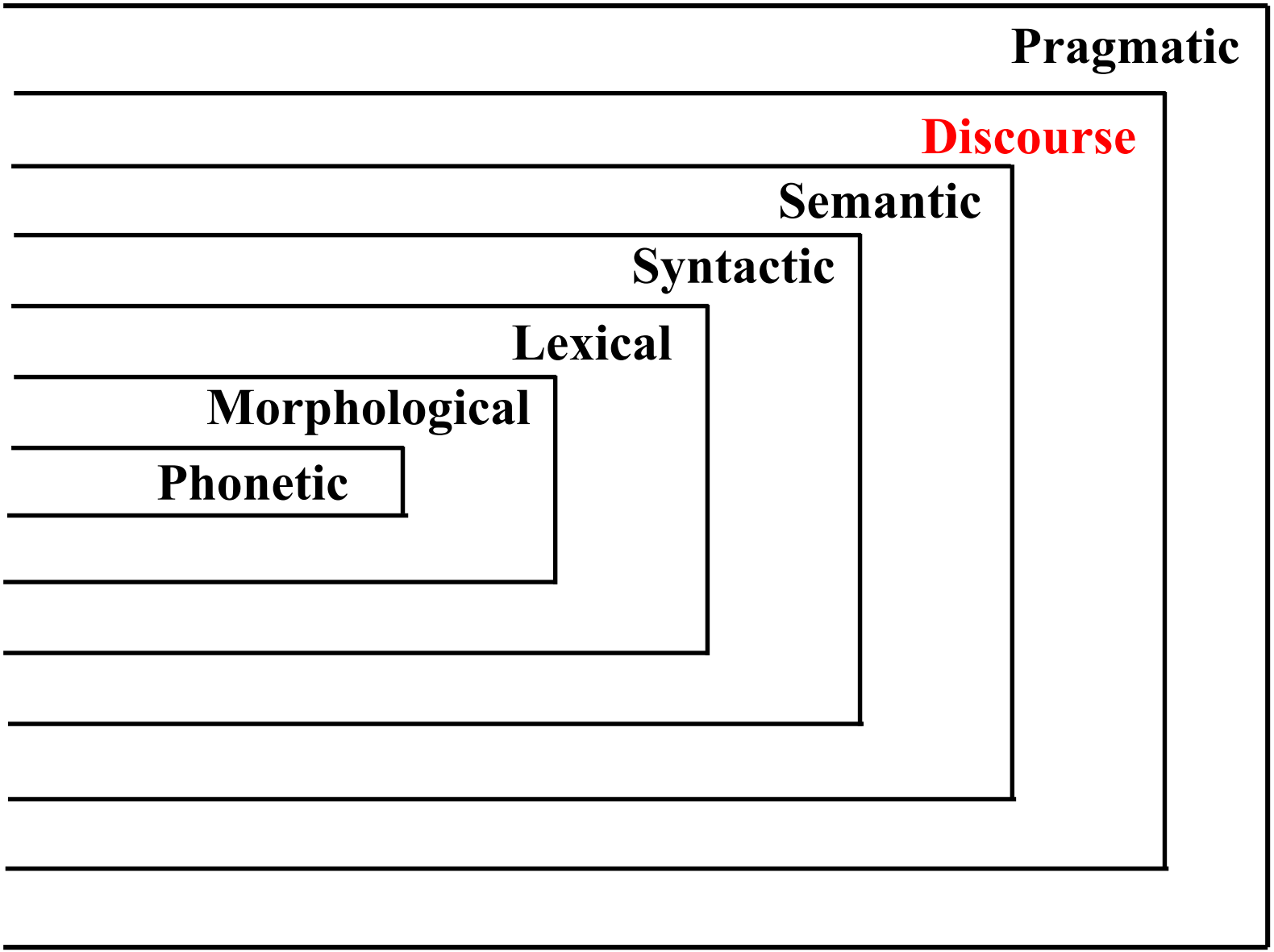
INTRODUCTION TO DISCOURSE LINGUISTICS AND DISCOURSE STRUCTURE

Lu Xiao
lxiao04@syr.edu
213 Hinds Hall

adopted some materials developed in previous courses by Nancy McCracken, Liz Liddy and others; and some instructor resources for the book "Speech and Language Processing" by Daniel Jurafsky and James H. Martin

SYNCHRONIC MODEL OF LANGUAGE



DISCOURSE LINGUISTICS

" *No one is in a position to write a comprehensive account of discourse analysis. The subject is at once too vast, and too lacking in focus and consensus*." (Stubbs, Discourse Analysis)

Definitional Elements

- Study of texts (linguistic units) larger than a sentence.
- Text is more than a sequence of sentences to be considered one by one.
  - Rather, sentences of a text are elements whose significance resides in the contribution they make to the development of a larger whole.
- Each type of text has its own structure that can convey meaning to the reader.
- Some issues of discourse understanding are closely related to those in pragmatics which studies the real world dependencies of utterances.

Distinctions Between Text And Discourse

- In some contexts, e.g. in communication research, the word discourse means
  - interactive conversation
  - spoken
- And the word text means
  - non-interactive monologue
  - Written
- But for (American) linguists, the word discourse can mean both of these things at the discourse level.

Scope of Discourse Analysis

- What does discourse analysis extract from text more than the explicit information discoverable by sentence-level syntax and semantics methodologies?
  - Structural organization of the text
  - Overall topic(s) of the text
  - Features which provide *cohesion* to the text
- What linguistic features of texts reveal this information to the analyst?

Discourse Segmentation

- Documents are automatically separated into passages, sometimes called fragments, which are different discourse segments
  - Discourse segments can inform semantic interpretation of document
- Techniques to separate documents into passages include
  - Rule-based systems based on clue words and phrases
  - Probabilistic techniques to separate fragments and to identify discourse segments (Oddy)
  - Lexical cohesion to identify fragments (TextTiling)

Texttiling

- Uses lexical cohesion to identify segments, assuming that each segment exhibits "lexical cohesion" within the segment, but is not cohesive across different segments
- Algorithm
  - Identifies candidate segments
  - Computes lexical cohesion score in each segment – Lexical cohesion score is the average semantic similarity of words within a segment
  - Identify boundaries by the difference of cohesion scores
  - NLTK has a text tiling algorithm available

Discourse Structure

- Human discourse often exhibits structures that are intended to indicate common experiences and respond to them
  - For example, research abstracts are intended to inform readers in the same community as the authors and who are engaged in similar work
  - Essay structure taught to high school students
  - Newspaper structure, where the story is given in several segments lending itself to shorter or longer versions

Discourse Relations

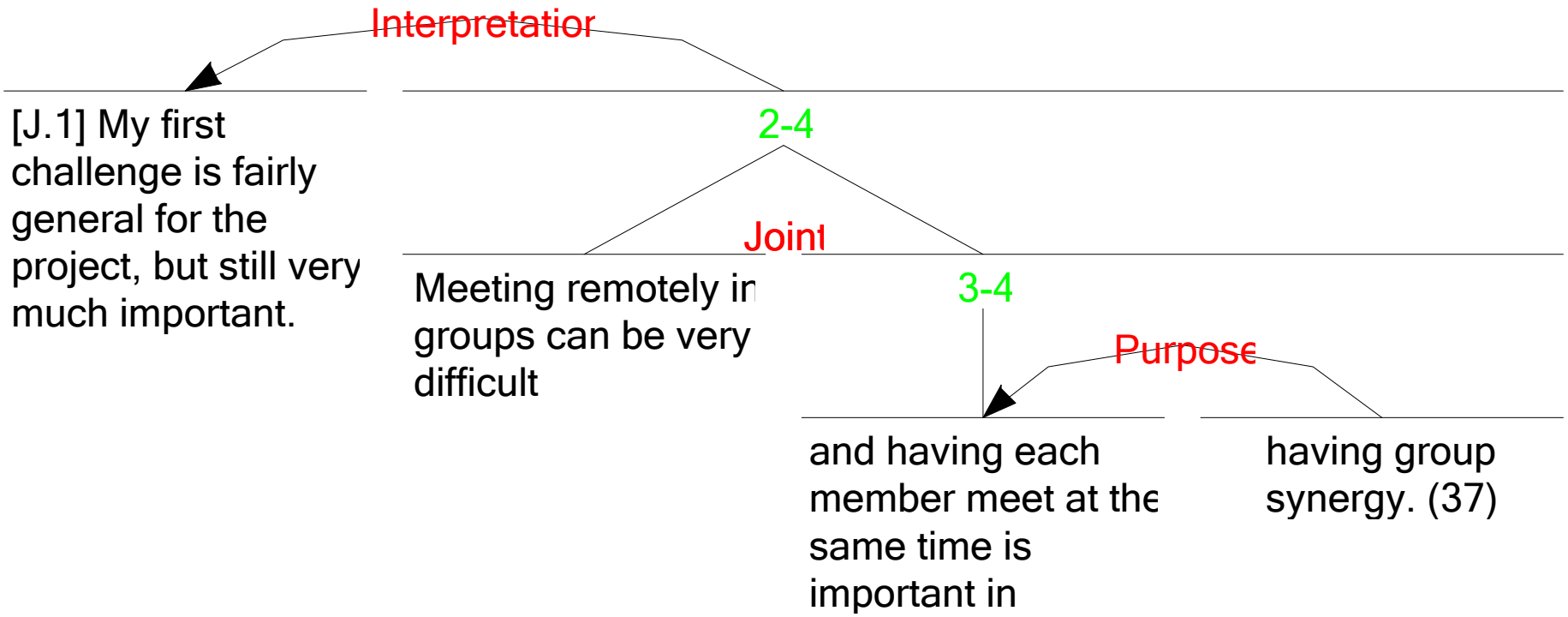
Rhetorical Structure Theory (RST): a theory of text organization created in the 1980s

Text units as *nuclear* and *satellites*

Three categories of relations:
subject matter relations,
presentational relations,
multinuclear relations

http://www.sfu.ca/rst/

Rhetorical Structure Theory



RST Annotation In The Rationale Texts (Xiao & Conroy, 2017)

Rutgers Argument Mining Corpora
(Wacholder, Muresan, Ghosh, Aakhus, 2014)

Re: #18 You must be a shill for the RIAA and others. Truth is that file sharing in fact has been a boon to truly independent filmmakers and other non affiliated content producers, because they've become better known; it only really harms the big sharks, who have been robbing everyone blind and suppressing the little guy for decades.

Secondly, piracey is illegal. You aren't sharing anything. It's just a eufimisom for STEALING, and that's exactly what it is. It cheats the creator out of their money.

RST Annotation In The Rationale Texts (Xiao & Conroy, 2017)

Common RST relations in the callouts that have rationales

"Common" RST relations	Percentage in Coded RST relations
Justify	19.11%
Elaboration	13.31%
Conjunction	9.74%
Contrast	8.81%
Joint	7.73%
Evaluation	5.93%
Circumstance	4.75%
Condition	4.61%
Non Volitional Cause	3.81%
Concession	3.22%
	81.03%

Corpora	Percentage in coded RST relations
android	77.81%
ban	76.79%
ipad	81.82%
layoff	84.08%
twitter	82.05%

Computational Analysis: Discourse Parsing

- Discourse Segmentation
- Discourse Relation Detection
- Rhetorical Tree Building

Li, J., Li, R., & Hovy, E. H. (2014). Recursive Deep Models for Discourse Parsing. In *EMNLP* (pp. 2061-2069)

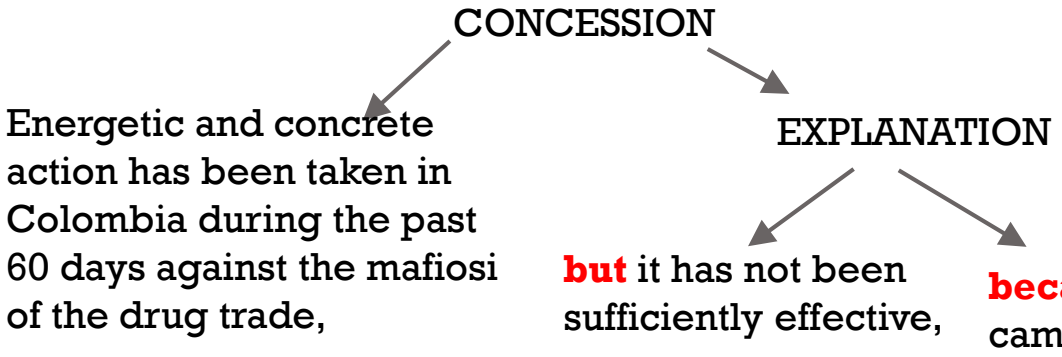
da Cunha, I. (2013). A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers. *Research in Computing Science*, *70*, 95-106.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

Rationale Detection: A Corpus-based Lexical Cue Graph Model (Khazaei & Xiao, 2015; Khazaei, Xiao, & Mercer, 2015)

Step 1: Identify the **RST relations that are commonly present** in the rationale texts

Step 2: Identify the **lexical cues for these RST relations** based on **two corpora**



Corpora

- RST Discourse Treebank
- SFU Review Corpus

Explicit relations are the ones that are signaled by cues:

- ***lexical cues***
- punctuations
- mood
- ..

Corpora

- RST Corpus - News corpus
  - 385 Wall Street Journal Articles
- SFU Corpus – Reviews corpus
  - 400 reviews from movie, book, and products

Lexical Cue Extraction

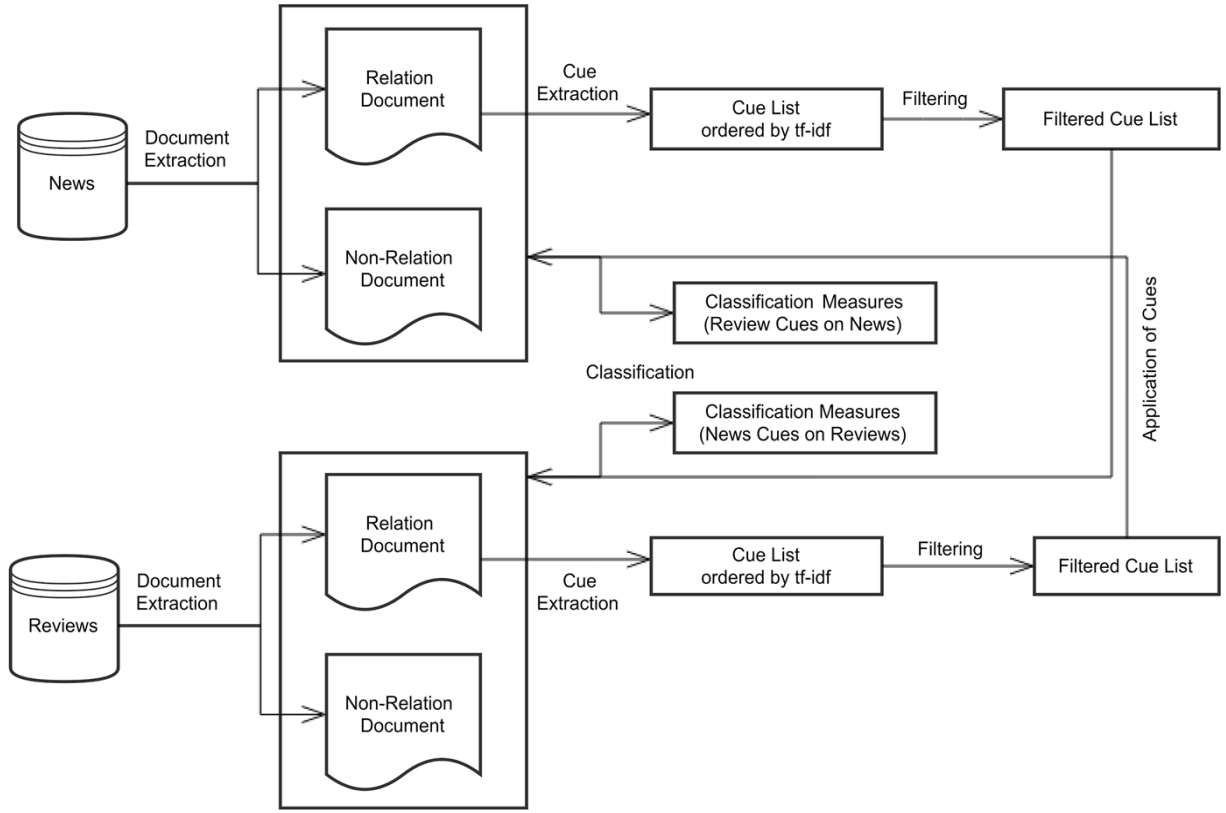
Biran And Rambow, 2011

- Extract all of the relation instances from the corpus and form a document named after the relation
- Extract all the n-grams from the document
- Calculate TF-IDF, sort the n-grams, and filter the list (to exclude any pronoun, or auxiliary verb)

Example:

Relation	**News**	**Reviews**
CIRCUMSTANCE	when, now, since	while, until, once
EVALUATION	good, high, well	nice, love it, impress
ELABORATION	who, which, as	which, where, as if

Experiment Description



Finding – CIRCUMSTANCE, ELABORATION, EVALUATION

- CIRCUMSTANCE
  - Heavily signaled
  - Signals are common discourse markers
  - Relatively genre-independent
- EVALUATION
  - Cues are not traditional markers
  - Dependent on the underlying genre, may or may not be signaled
- ELABORATION
  - Low weight and F-score
  - Corpus-based lexical cue approach is NOT recommended

Discourse Linguistics: Text Cohesion And Coherence

Cohesion and Coherence

- A text will exhibit unity / texture
  - on the surface level (cohesion)
  - at the meaning level (coherence)
- Halliday & Hasan's Cohesion in English (1976)
  - Categories of the surface level ties
  - Sets forth the linguistic devices that are available in the English language for creating this unity / texture
  - Identifies the features in a text that contribute to an intelligent comprehension of the text

Cohesive Relations

- Define dependencies between sentences in text.

"He said so."

"*He*" and "*so*" presuppose elements in the preceding text for their understanding

- This presupposition and the presence of information elsewhere in text to resolve this presupposition provide COHESION

- Part of the discourse-forming component of the linguistic system

- Provides the means whereby structurally unrelated elements are linked together

Six Types Of Cohesive Ties

- Grammatical
  - Reference
  - Substitution
  - Ellipsis
  - Conjunction
- Lexical
  - Reiteration
  - Collocation
- (In practice, there is overlap; some examples can show more than one type of cohesion.)

1. Reference

- items in a language which, rather than being interpreted in their own right, make reference to something else for their interpretation.

- Anaphora example with references (he, his, there) to previous text:

  "Doctor Foster went to Gloucester in a shower of rain. He stepped in a puddle right up to his middle and never went there again."

- Cataphora example with pronoun (he) referring to following text.

  When he visited the construction site last month, Mr. Jones talked with the union leaders about their safety concerns.

2. **Substitution**:

- a substituted item that serves the same structural function as the item for which it is substituted.

  Nominal – *one, ones, same*

  Verbal – *do*

  Clausal – *so, not*

  - *These biscuits are stale. Get some fresh ones.*
  - Person 1 – *I'll have two poached eggs on toast, please.*

    Person 2 – *I'll have the same.*

  - *The words did not come the same as they used to do. I don't know the meaning of half those long words, and what's more, don't believe you do either, said Alice.*

3. Ellipsis

- Very similar to substitution principles, embody same relation between parts of a text
- Something is left unsaid, but understood nonetheless, but a limited subset of these instances
  - *Smith was the first person to leave. I was the second* _____.
  - *Joan brought some carnations and Catherine* _____ *some sweet peas.*
  - *Who is responsible for sales in the Northeast? I believe Peter Martin is* _____.

4. Conjunction

- Different kind of cohesive relation in that it doesn't require us to understand some other part of the text to understand the meaning

- Rather, a specification of the way the text that follows is systematically connected to what has preceded

For the whole day he climbed up the steep mountainside, almost without stopping.

And in all this time he met no one.

Yet he was hardly aware of being tired.

So by night the valley was far below him.

Then, as dusk fell, he sat down to rest.

Now, 2 types of Lexical Cohesion

Lexical cohesion is concerned with cohesive effects achieved by selection of vocabulary

5. Reiteration continuum –

I attempted an ascent of the peak. _X__ was easy.

- same lexical item – *the ascent*
- synonym – *the climb*
- super-ordinate term – *the task*
- general noun – *the act*
- pronoun - *it*

6. Collocations

Lexical cohesion achieved through the association of semantically related lexical items

- Accounts for any pair of lexical items that exist in some lexico-semantic relationship, e. g.

- complementaries
  boy / girl
  stand-up / sit-down
- antonyms
  wet / dry
  crowded / deserted
- converses
  order / obey
  give / take
- pairs from ordered series
  Tuesday / Thursday
  sunrise / sunset
- part-whole
  brake / car
  lid / box
- co-hyponyms of same super-ordinate
  chair / table (furniture)
  walk / drive (go)

Uses Of Cohesion Theory

- Scoring text cohesiveness
  - Halliday & Hasan's theory has been captured in a coding scheme used to quantitatively measure the extent of cohesion in a text.
  - ETS has experimented with it as a metric in grading standardized test essays.

- Language generation and machine translation can use cohesion and coherence to build fluent texts

Building Semantic Representations

- When building a semantic representation of a text, the theory suggests how the system can recognize relations between entities.
  - Which entities in the text are related
  - How they are related
    - Particularly, coreference resolution finds all of the references to the "same" entity and groups them into clusters
- Information Extraction requires coreference resolution to build the relation triples

Lexical Chains

- Building lexical chains is one way to find the lexical cohesion structure of a text, both reiteration and collocation.
- A lexical chain is a sequence of semantically related words from the text
- Document can be viewed as a set of lexical chains
  - A kind of clustering of words based on semantic similarity
  - Each cluster can be viewed as a document "topic"
- Algorithm sketch:
  - Select a set of candidate words
  - For each candidate word, find an appropriate chain relying on a "relatedness" measure among members of chains
    - Usually semantic similarity between words
  - If it is found, insert the word into the chain.

Coherence Relations – Semantic Meaning Ties

- The set of possible relations between the meanings of different utterances in the text
- Hobbs (1979) suggests relations such as
  - **Result:** state in first sentence could cause the state in a second sentence
  - **Explanation:** ,the state in the second sentence could cause the first
    John hid Bill's car keys. He was drunk.
  - **Parallel:** The states asserted by two sentences are similar
    The Scarecrow wanted some brains. The Tin Woodsman wanted a heart.
  - **Elaboration:** Infer the same assertion from the two sentences.
- Textual Entailment
  - NLP task to discover the result and elaboration between two sentences
- The examination of transitional phrases and the measure of lexical overlap in our study (Khazaei, Xiao, & Mercer , 2017)

Khazaei, T., Xiao, L., & Mercer, R. (2017). Writing to Persuade: Analysis and Detection of Persuasive Discourse