
Discourse Linguistics: Text Coherence and Cohesion

Cohesion and Coherence

- “A piece of text is intended and is perceived as more than a simple sequencing of independent sentences.”
- Therefore, 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. (This is far less common.)

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

* Resolving these references is the **coreference task**, which is one of the most important NLP tasks at the discourse level.

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, I don't believe you do either, said Alice.*

* These examples will also be treated by the **coreference task** (except for verbs)

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*

* These are similar to substitution and will also be treated by **coreference**.

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.

Example of Lexical Chains

- Semantically related words are highlighted. Note that this assumes that coreference resolution has already occurred.

From Nineteen Eighty-Four [abridged]:

... the **book** that he had just taken out of the drawer. **It** was a peculiarly beautiful **book**. **Its** smooth creamy **paper** was of a kind that had not been manufactured for at least forty years past. He could guess, however, that the **book** was much older than that. He had seen **it** lying in the window of a frowsy little **junk-shop** and had been stricken immediately by an overwhelming desire to **possess** **it**. Party members were supposed not to go into ordinary **shops**. He had slipped inside and **bought** the **book** for two **dollars** fifty. Even with nothing **written** in **it**, **it** was a compromising **possession**. ¶ The thing that he was about to do was to open a **diary**. Winston fitted a **nib** into the **penholder** and sucked **it** to get the grease off. The **pen** was an archaic **instrument**, seldom used even for **signatures**, and he had **procured** one, furtively and with some difficulty, simply because of a feeling that the beautiful creamy **paper** deserved to be **written** on with a real **nib** instead of being scratched with an **ink**-**pencil**. Actually he was not used to **writing** by hand. He dipped the **pen** into the **ink**.

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.