Discourse Linguistics:
  Discourse Structure
  Text Coherence and Cohesion
  Reference Resolution
Synchronic Model of Language

- Phonetic
- Morphological
- Lexical
- Syntactic
- Semantic
- Discourse
- Pragmatic
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.

- Texts have their own structure and way of conveying meaning.

- Some issues of discourse understanding are closely related to those in pragmatics which studies the real world dependence of utterances.
Distinctions Between Text and Discourse

- In some contexts, 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.
Features which Characterize Spoken Language

1. Less structured syntax, including many incomplete sentences.
2. Little subordination.
3. Preponderance of active declarative forms.
4. Clauses are conjoined by conjunctives such as and, but, then, rather than more formal ones such as firstly, more importantly, in conclusion.
5. Shorter noun phrases with many fewer premodifiers.
6. Speaker continues to refine expressions (this man, Uh, this fellow she was going out with).
7. A good deal of rather generalized vocabulary, such as a lot of, got to, thing, do, stuff.
8. Many fillers, such as well, ahem, I think, you know, of course.
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 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

• Empirical study in dissertation by Liz Liddy identifies discourse structure of research abstracts
  – Hierarchical, componential text structure
Discourse Segmentation

• Documents are automatically separated into passages, sometimes called fragments, which are different discourse segments

• Techniques include
  – Rule-based systems based on clue words and phrases
  – Probabilistic techniques to separate fragments and to identify discourse segments (Oddy)
  – TextTiling algorithm uses cohesion to identify segments, assuming that each segment exhibits lexical cohesion within the segment, but is not cohesive across different segments
    • Lexical cohesion score – average similarity of words within a segment
    • Boundary identification
Cohesion

- “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)
  - 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
  - For generation, produces natural-sounding texts
Cohesive Relations

- Exist between elements in a text where the interpretation of one is dependent on others

  “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.

“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.”

Types of Reference

- exophora
  - situation – referring to things outside of text – not part of cohesion

- endophora
  - textual

- anaphora
  - preceding text

- cataphora
  - following text
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

  - An elliptical item is one which leaves specific structural slots to be filled from elsewhere else

- As compared to substitution, where a place-marker is inserted in the structural slot

  - In ellipsis, substitution by zero.

- Types

  - Nominal, Verbal, Clausal
Ellipsis Examples

• *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

- 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
Collocations (cont’d)

- 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

1. 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.

2. When building a semantic representation of a text, the theory suggests how the system can recognize relations between entities.
   - indicates **what** is related
   - suggests **how** they are related

3. Provides guidance to a NL Generation system so that the system can produce naturally cohesive text.

4. Delineates (for English) how the cohesive features of the language can be recognized and utilized by an MT system.
Coherence Relations

- 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.
Anaphora / Reference Resolution

- A linguistic phenomenon of abbreviated subsequent reference
  - A cohesive tie of the grammatical and lexical types
    - Includes reference, substitution and reiteration

  - A technique for referring back to an entity which has been introduced with more fully descriptive phrasing earlier in the text

  - Refers to this same entity but with a lexically and semantically attenuated form
Types of Entity Resolutions

- **Entity Resolution** is an ability of a system to recognize and unify variant references to a single entity.

- 2 levels of resolution:
  - within document (**co-reference resolution**)
    - e.g. *Bin Ladin* = *he*
    - *his followers* = *they*
    - *terrorist attacks* = *they*
    - *the Federal Bureau of Investigation* = *FBI* = *F.B.I*
  - across document (**named entity resolution**)
    - e.g. *maverick Saudi Arabian multimillionaire* = *Usama Bin Ladin* = *Bin Ladin*

- **Event resolution** is also possible, but not widely used
1. **The State Department** renewed its appeal for **Bin Laden** on Monday and warned of possible fresh attacks by his followers against U.S. targets.

2. One early target of the F.B.I.’s Budapest office is expected to be **Semyon Y. Mogilevich, a Russian citizen who** has operated out of Budapest for a decade. Recently he has been linked to the growing money-laundering investigation in the United States involving the Bank of New York. **Mr. Mogilevich** is also the target of a separate money laundering and financial fraud investigation by the F.B.I. in Philadelphia, according to federal officials.

3. **The F.B.I.** will also have the final say over the hiring and firing of the 10 Hungarian agents who will work in the office, alongside five American agents. **The bureau** has long had agents posted in American embassies.
Glossary of Terminology

• **Referring phrase = Anaphora = Anaphoric Expression = Co-reference = Coreference**
  – an expression that identifies an earlier mentioned entity (including pronouns and definite noun phrases)

• **Referent = Antecedents** entity that a referring phrase refers back to

• **Referent Candidates** - all potential entities / antecedents that a referring phrase could refer to

• **Alias = Named Entity** - a cross document co-reference
  – includes proper names (mostly)
Terminology Examples

Referent Candidates for “the victim”

- Unidentified gunmen shot dead a businessman in the Siberian town of Leninsk-Kuznetsk on Wednesday, but the victim was not linked to the Sibneft oil major as originally thought, police and company officials said. (afp19980610.1.sgm). He appears to be associated with local …

Referring phrases
Reference Types
Definite noun phrases – the X

• Definite reference is used to refer to an entity identifiable by the reader because it is either
  – a) already mentioned previously (in discourse), or
  – b) contained in the reader’s set of beliefs about the world (pragmatics), or
  – c) the object itself is unique. (Jurafsky & Martin, 2000)

• E.g.
  – Mr. Torres and his companion claimed a hardshelled black vinyl suitcase. The police rushed the suitcase (a) to the Trans-Uranium Institute (c) where experts cut it open because they did not have the combination to the locks.

  – The German authorities (b) said a Colombian who had lived for a long time in the Ukraine (c) flew in from Kiev. He had 300 grams of plutonium in his baggage. The suspected smuggler (a) denied that the materials (a) were his.
Pronominalization

- **Pronouns** refer to entities that were introduced fairly recently, 1-4-5-10(?) sentences back.
  - *Nominative* (he, she, it, they, etc.)
    - e.g. The German authorities said a Colombian who had lived for a long time in the Ukraine flew in from Kiev. He had 300 grams of plutonium 239 in his baggage.
  - *Oblique* (him, her, them, etc.)
    - e.g. Undercover investigators negotiated with three members of a criminal group and arrested them after receiving the first shipment.
  - *Possessive* (his, her, their, etc. + hers, theirs, etc.)
    - e.g. He had 300 grams of plutonium 239 in his baggage. The suspected smuggler denied that the materials were his. (*chain)
  - *Reflexive* (himself, themselves, etc.)
    - e.g. There appears to be a growing problem of disaffected loners who cut themselves off from all groups.
Indefinite noun phrases – a X, or an X

• Typically, an indefinite noun phrase introduces a new entity into the discourse and would not be used as a referring phrase to something else
  – The exception is in the case of cataphora:
    *A Soviet pop star was killed at a concert in Moscow last night.* *Igor Talkov was shot through the heart as he walked on stage.*
  – Note that cataphora can occur with pronouns as well:
    *When he visited the construction site last month, Mr. Jones talked with the union leaders about their safety concerns.*
Demonstratives – this and that

- Demonstrative pronouns can either appear alone or as determiners
  
  *this ingredient, that spice*

- These NP phrases with determiners are ambiguous
  - They can be indefinite
    
    *I saw this beautiful car today.*
  - Or they can be definite
    
    *I just bought a copy of Thoreau’s Walden. I had bought one five years ago. That one had been very tattered; this one was in much better condition.*
Names

• Names can occur in many forms, sometimes called name variants.

Victoria Chen, Chief Financial Officer of Megabucks Banking Corp. since 2004, saw her pay jump 20% as the 37-year-old also became the Denver-based financial-services company’s president. Megabucks expanded recently. . . . MBC. . .

  – (Victoria Chen, Chief Financial Officer, her, the 37-year-old, the Denver-based financial-services company’s president)
  – (Megabucks Banking Corp. , the Denver-based financial-services company, Megabucks, MBC )

• Groups of a referent with its referring phrases are called a coreference group.
Unusual Cases

• Compound phrases
  
  *John and Mary got engaged. They make a cute couple.*
  
  *John and Mary went home. She was tired.*

• Singular nouns with a plural meaning

  *The focus group met for several hours. They were very intent.*

• Part/whole relationships

  *John bought a new car. A door was dented.*

*Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.*
Approach to coreference resolution

• Naively identify all referring phrases for resolution:
  – all Pronouns
  – all definite NPs
  – all Proper Nouns

• Filter things that look referential but, in fact, are not
  – e.g. geographic names, the United State
  – pleonastic “it”, e.g. it’s 3:45 p.m., it was cold
  – non-referential “it”, “they”, “there”
    • e.g. it was essential, important, is understood,
    • they say,
    • there seems to be a mistake
All noun phrases (both indef. and def.) are considered potential referent candidates.

A referring phrase can also be a referent for a subsequent referring phrases,

- Example: (omitted sentence with name of suspect)
  He had 300 grams of plutonium 239 in his baggage. The suspected smuggler denied that the materials were his.
  (chain of 4 referring phrases)

All potential candidates are collected in a table collecting feature info on each candidate.

Problems:
- chunking
  - e.g. the Chase Manhattan Bank of New York
- nesting of NPs
Features

• Define features between a referring phrase and each candidate
  – Number agreement: plural, singular or neutral
    • He, she, it, etc. are singular, while we, us, they, them, etc. are plural and should match with singular or plural nouns, respectively
    • Exceptions: some plural or group nouns can be referred to by either it or they
      IBM announced a new product. They have been working on it ...
  – Gender agreement:
    • Generally animate objects are referred to by either male pronouns (he, his) or female pronouns (she, hers)
    • Inanimate objects take neutral (it) gender
  – Person agreement:
    • First and second person pronouns are “I” and “you”
    • Third person pronouns must be used with nouns
More Features

• Binding constraints
  – Reflexive pronouns (himself, themselves) have constraints on which nouns in the same sentence can be referred to:
    \[ \text{John bought himself a new Ford. (John = himself)} \]
    \[ \text{John bought him a new Ford. (John cannot = him)} \]

• Recency
  – Entities situated closer to the referring phrase tend to be more salient than those further away
    • And pronouns can’t go more than a few sentences away

• Grammatical role / Hobbs distance
  – Entities in a subject position are more likely than in the object position
Even more features

- Repeated mention
  - Entities that have been the focus of the discourse are more likely to be salient for a referring phrase

- Parallelism
  - There are strong preferences introduced by parallel constructs
    \textit{Long John Silver went with Jim. Billy Bones went with him.}
    \textit{(him = Jim)}

- Verb Semantics and selectional restrictions
  - Certain verbs take certain types of arguments and may prejudice the resolution of pronouns
    \textit{John parked his car in the garage after driving it around for hours.}
Example: rules to assign gender info

- Assign gender to “masculine”,
  - if it is a pronoun “he, his, him”
  - if it contains markers like “Mr.”
  - if the first name belongs to a list of masculine names

- Same for “feminine” and “neuter” (except for latter use categories such as singular, geo names, company names, etc.)

- Else, assign “unknown”
Approaches

• Assign weights to the features and resolve the referring phrase to the candidate which achieves the highest score by summing over the weighted features

• Train a classifier over an annotated corpus to identify which candidates and referring phrases are in the same coreference group
  – Evaluation results (for example, Vincent Ng at ACL 2005) are on the order of F-measure of 70, with generally higher precision than recall
  – Evaluation typically uses the B-Cubed scorer introduced by Bagga and Baldwin, which compares coreference groups
  – Pronoun coreference resolution by itself is much higher scoring, usually over 90%. 