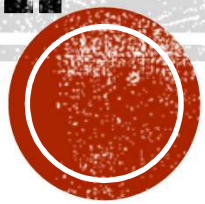
Discourse Linguistics: Coreference Resolution



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

Anaphora/Co-Reference Resolution

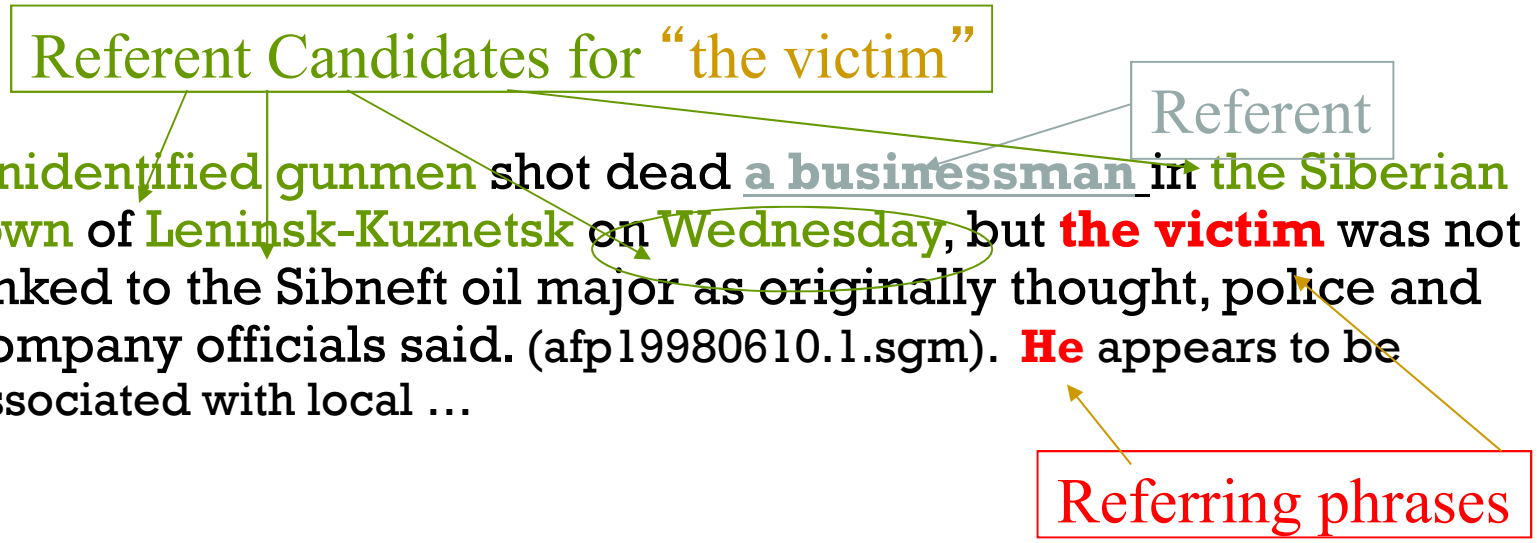
- A linguistic phenomenon of abbreviated subsequent reference
  - A cohesive tie of the grammatical and lexical types
    - Includes reference, substitution and reiteration
  - A technique for referring back to an entity which has been introduced with more fully descriptive phrasing earlier in the text
  - 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.
  - Coreference algorithms usually performed within larger task of entity resolution
- 2 levels of resolution:
  - within document (includes **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 (or **named entity resolution**)
    - e.g. *Usama Bin Ladin =Osama Bin Ladin = Bin Ladin*
- **Event resolution** is also possible
  - Bejan, C. A., & Harabagiu, S. (2014). Unsupervised event coreference resolution. *Computational Linguistics*, *40*(2), 311-347.

Terminology Examples

- The referent for a referring phrase is found by the resolution algorithm among the candidates, previous noun phrases.



Reference Types

- An algorithm must first decide which are the referring phrases that must be resolved
  - Pronouns
  - Definite noun phrases (the)
  - Indefinite noun phrases (a, an)
  - Names
  - Others

Pronouns

- **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$_1$** who had lived for a long time in the Ukraine flew in from Kiev. **He$_1$** 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$_2$** and arrested **them$_2$** after receiving the first shipment.

Pronouns

- **Pronouns** refer to entities that were introduced fairly recently, 1-4-5-10(?) sentences back.
  - ***Possessive*** (his, her, their, etc. + hers, theirs, etc.)
    - e.g. **He$_3$** had 300 grams of plutonium 239 in **his$_3$** baggage. The suspected smuggler$_3$* denied that the materials were **his$_3$**. ***(*chain)***
  - ***Reflexive*** (himself, themselves, etc.)
    - e.g. There appears to be a growing problem of **disaffected loners$_4$** who cut **themselves$_4$** off from all groups .
  - ***Demonstrative pronouns*** (this, that, etc.) can either appear alone or as determiners

    this ingredient, that spice
    - These NP phrases with determiners are ambiguous: they can be indefinite (e.g., I saw this beautiful car today) or definite (e.g., 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).

Definite Noun Phrases – The X

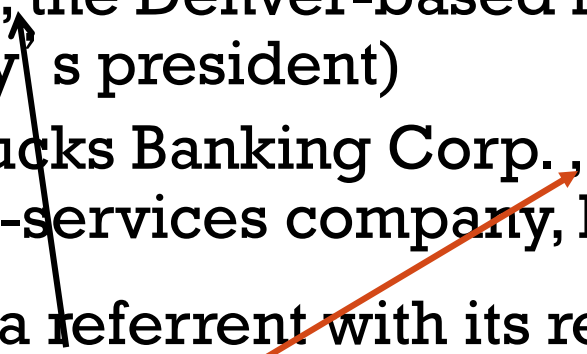
- Definite reference is used to refer to an entity identifiable by the reader because it is either
  - a) already mentioned previously (in discourse), or
  - b) contained in the reader's set of beliefs about the world (pragmatics), or
  - c) the object itself is unique. (Jurafsky & Martin, 2000)
- E.g.
  - Mr. Torres and his companion claimed **a hardshelled black vinyl suitcase$_1$**. The police rushed **the suitcase$_1$** (a) to **the Trans-Uranium Institute$_2$** (c) where experts cut **it$_1$** open because they did not have the combination to the locks.
  - **The German authorities$_3$** (b) said **a Colombian$_4$** who had lived for a long time in **the Ukraine$_5$** (c) flew in from Kiev. He had **300 grams of plutonium 239$_6$** in his baggage. **The suspected smuggler$_4$** (a) denied that **the materials$_6$** (a) were his.

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.

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 referrent 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.

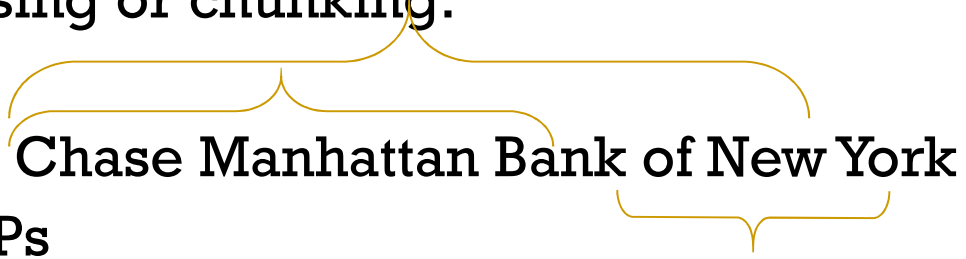
Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.

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 States*
  - 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*

Identify Referent Candidates

- 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.
- Requires either parsing or chunking:
  - chunking
    - e.g. the Chase Manhattan Bank of New York
  - Note nesting of NPs



Some 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

Some Features (cont'd)

- Binding constraints
  - Reflexive pronouns (himself, themselves) have constraints on which nouns in the same sentence can be referred to:
    John bought himself a new Ford. (John = himself)
- 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, sometimes approximated by Hobbs distance
  - Entities in a subject position are more likely than in the object position

Approaches

- 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 0.70, with generally higher precision than recall

- Lassalle, E., & Denis, P. (2015, January). Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures. In *AAAI* (pp. 2274-2280)
  - employ latent trees to represent the full coreference and anaphoricity structure of a document

- Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D. E., & Cudré-Mauroux, P. (2015, October). SANAPHOR: ontology-based coreference resolution. In *International Semantic Web Conference* (pp. 458-473). Springer International Publishing
  - Leverages Semantic Web technologies

Summary Of Discourse Level Tasks

- Dialogue structure (discourse segmentations, discourse relations, text coherence)
- Document structure
  - Recognizing known structure, for example, abstracts
  - Separating documents according to known structure
- Named entity resolution across documents

Topic Modelling

- Examination of patterns in the given text(or corpus) at the semantic level by extracting *topics* from *texts*
- What is a topic?
- A list of words that occur in statistically meaningful ways (for the computer)
- What is a text?
- Unstructured text such that no computer-readable annotations available that indicate the semantic meaning of the words in the text
- Is it always useful?
  - Small number of documents (or even a single document) – maybe not
  - Hundreds of documents – may be

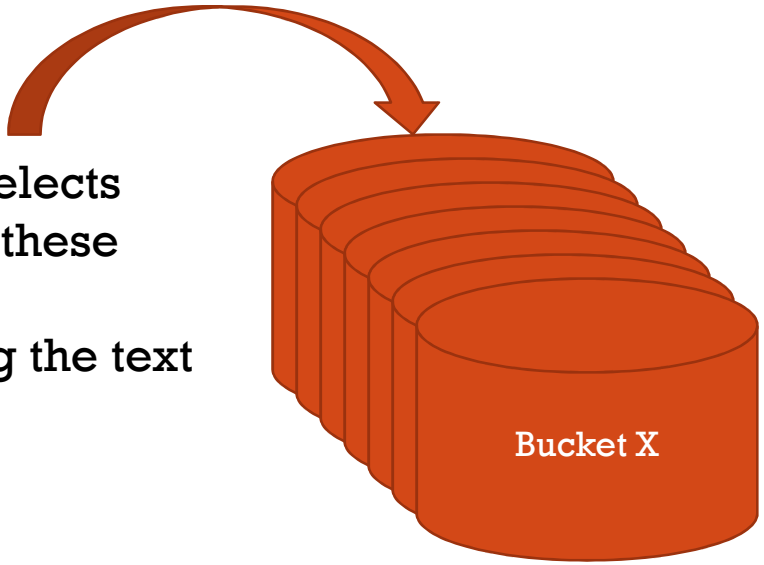
Topic Modelling - LDA

- An unsupervised method that models documents and topics based on Dirichlet distribution
  - each document is considered to be a distribution over various topics and each topic is modeled as a distribution over words.
- To model the distributions, LDA also requires the number of topics (often denoted by *k*) as an input.

  For instance, the following are the topics extracted from a random set of tweets from Canadian users where k = 3:
  - Topic 1: great, day, happy, weekend, tonight, positive experiences
  - Topic 2: food, wine, beer, lunch, delicious, dining
  - Topic 3: home, real estate, house, tips, mortgage, real estate

LDA Topic Modeling: Assumption

An author selects words from these buckets in constructing the text



A bucket contains words that are related to one topic

Topic Modeling - LDA

	W1	W2	W3	…	Wm
D1					
D2					
D3					
…					
Dn					

P1: proportion of words in document d that are currently assigned to topic t

P2: proportion of assignments to topic t over all documents that come from word w

	K1	K2	K3	…	K
D1					
D2					
D3					
…					
Dn					

	W1	W2	W3	…	Wm
K1					
K2					
K3					
…					
K					

Topic Modeling - LDA

- Domain knowledge
  - α:
    - a low α value: more weight on having each document being composed of only a few dominant topics,
    - a high α value: more weight on having each document being composed of a relatively larger set of topics.
  - β:
    - a low β value: more weight on having each topic being composed of only a few dominant words.
    - a high β value: more weight on having each topic being composed of a relatively larger set of topics.
- Number of topics: Kullback Leibler Divergence Score (KL divergence)
  - On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations

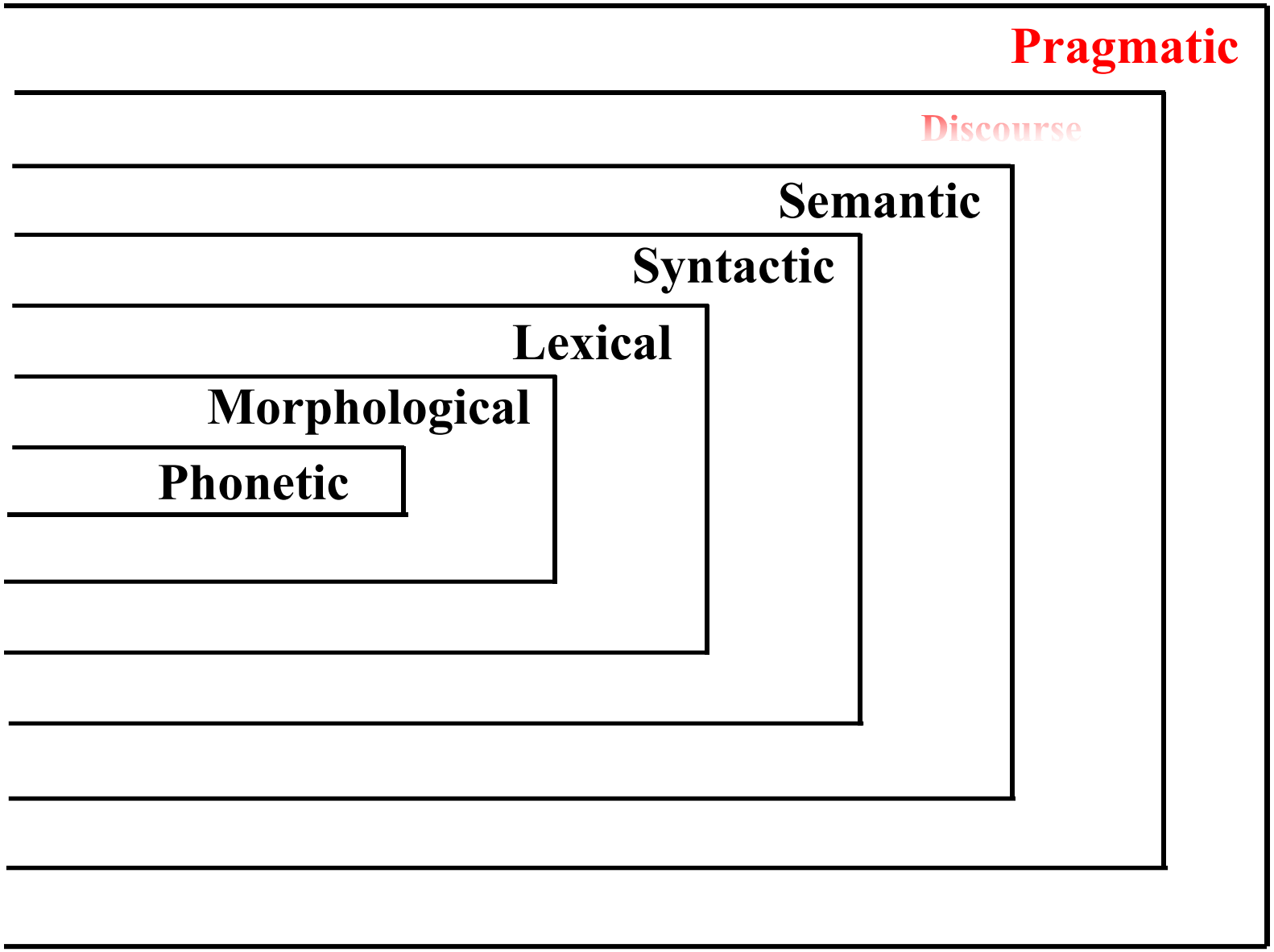
Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Topic Modeling

- Youtube video: https://www.youtube.com/watch?v=yK7nN3FcgUs
- Complete guide to topic modeling
  - https://nlpforhackers.io/topic-modeling/
- LDA (Latent Dirichlet Allocation) paper
  - http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Topic modeling in Python
  - Non-negative Matrix Factorization (NMF)
    - https://de.dariah.eu/tatom/topic_model_python.html
  - Gensim package

Pragmatics Level: Dialogue Theory

Synchronic Model Of Language



Pragmatics

- Functional perspective - The study of language in use
- Generally, aspects of language which require context to be understood
  - How the situational context is grammaticalized
  - World knowledge (knowledge bases) used for understanding
    - Useful pragmatics for semantic understanding of any text
- **One specific goal is to explain how extra meaning is read into utterance without actually being encoded in them**

Topics

- Theory: Properties of Human Conversations
  - **Speech Act Theory**
  - Gricean Maxims
  - Conversational Structure
  - Dialogue Act Theory

Speech Act Theory

A speech act is a performative utterance in communication

- Communication succeeds only if the intention of the speaker is recognized by the listener
  - Proposed by John Austin in 1962 in *How To Do Things With Words*
  - Systematized by John Searle in 1969 in *Speech Acts: An Essay in the Philosophy of Language*

- Propositional content (the literal meaning of the text) does not always fully communicate the speaker's intent
  - Example: *I'm going to pay you back for that.*

Speech Act Theory

- Three Levels of Speech Acts affecting the social reality of the speaker and listener:
  - Locutionary – proposition of speech act
    - The meaning of the sentence (what we've been working on in NLP)
  - Illocutionary – intention of speech act
    - The act of asking, answering, promising, etc. in uttering a sentence
  - Perlocutionary – consequences of speech act
    - The (often intentional) production of certain effects upon the feelings, thoughts, or actions of the addressee in uttering a sentence

"almost any speech act is really the performance of several acts at once, distinguished by different aspects of the speaker's intention: there is the act of saying something, what one does in saying it, such as requesting or promising, and how one is trying to affect one's audience." *Bach*

Speech Act Theory

- **Examples:**

"I'm telling you not to do that."

→Locutionary analysis: the speaker asks the listener not to do **that**

→Illocutionary analysis: this message is a ***warning***

→Perlocutionary analysis: if the listener does not do it after this message, the message is ***persuasive***

Speech Act Theory

- Three Levels of Speech Acts affecting the social reality of the speaker and listener:
  - Locutionary – proposition of speech act
    - The meaning of the sentence (what we've been working on in NLP)
  - **Illocutionary – intention of speech act**
    - **The act of asking, answering, promising, etc. in uttering a sentence**
  - Perlocutionary – consequences of speech act
    - The (often intentional) production of certain effects upon the feelings, thoughts, or actions of the addressee in uttering a sentence

"almost any speech act is really the performance of several acts at once, distinguished by different aspects of the speaker's intention: there is the act of saying something, what one does in saying it, such as requesting or promising, and how one is trying to affect one's audience." Bach

Taxonomy Of Illocutionary Acts

1. **Assertives** – commit the speaker to something's being the case – *suggest, swear, boast, conclude*

2. **Directives** – attempts by speaker to get listener to do something – *ask, order, request, invite, advise*

3. **Commissives** – obligate oneself to future course of action – *promise, plan, vow, oppose*

4. **Expressives** – share psychological state of speaker about something – *apologize, deplore, thank*

5. **Declarations** – bring about a different state of the world as a result of the utterance – *resign, baptize, marry*

Types Of Directives (Ervin-tripp, 1976)

Type of Directives	Example
Need statements	I need a match.
Imperatives	Give me a match.
Embedded imperatives	Could you give me a match?
Permission directives	May I have a match?
Question directives	Got a match?
Hints	The matches are all gone.

Imperatives

- Imperatives can also be used to express a suggestion, an invitation, a wish, an apology, etc.
  - Let's have dinner together. (suggestion)
  - Come in and have a seat. (invitation)
  - Have a good vacation! (wish)
  - Pardon me (apology)

Speech Act Theory

- Three Levels of Speech Acts affecting the social reality of the speaker and listener:
  - Locutionary – proposition of speech act
    - The meaning of the sentence (what we've been working on in NLP)
  - **Illocutionary – intention of speech act**
    - **The act of asking, answering, promising, etc. in uttering a sentence**
  - Perlocutionary – consequences of speech act
    - The (often intentional) production of certain effects upon the feelings, thoughts, or actions of the addressee in uttering a sentence

"almost any speech act is really the performance of several acts at once, distinguished by different aspects of the speaker's intention: there is the act of saying something, what one does in saying it, such as requesting or promising, and how one is trying to affect one's audience." Bach