Discourse Coherence and Review

COMP-599

Nov 5, 2015

Outline

- Cohesion and Coherence
- Theories of coherence
    - Rhetorical Structure Theory
    - Local coherence modelling
- Review for Midterm

Discourse

Language does not occur one sentence or utterance at a time.

Types of discourse:

Monologue – one-directional flow of communication

Dialogue – multiple participants

- Turn taking
- More varied communicative acts: asking and answering questions, making corrections, disagreements, etc.
- May touch on **human-computer interaction** (**HCI**)

Coherence

A property of a discourse that “makes sense” – there is some logical structure or meaning in the discourse that causes it to hang together.

Coherent:

Indoor climbing is a good form of exercise.

It gives you a whole-body workout.

Incoherent:

Indoor climbing is a good form of exercise.

Rabbits are cute and fluffy.

Cohesion

The use of linguistic devices to tie together text units

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Lexical Cohesion

Related words in a passage

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Coreference Chains

Anaphoric devices

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Discourse Markers

Cue words mark discourse relations

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must **also** inform a citizen that a stop is voluntary **and** they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

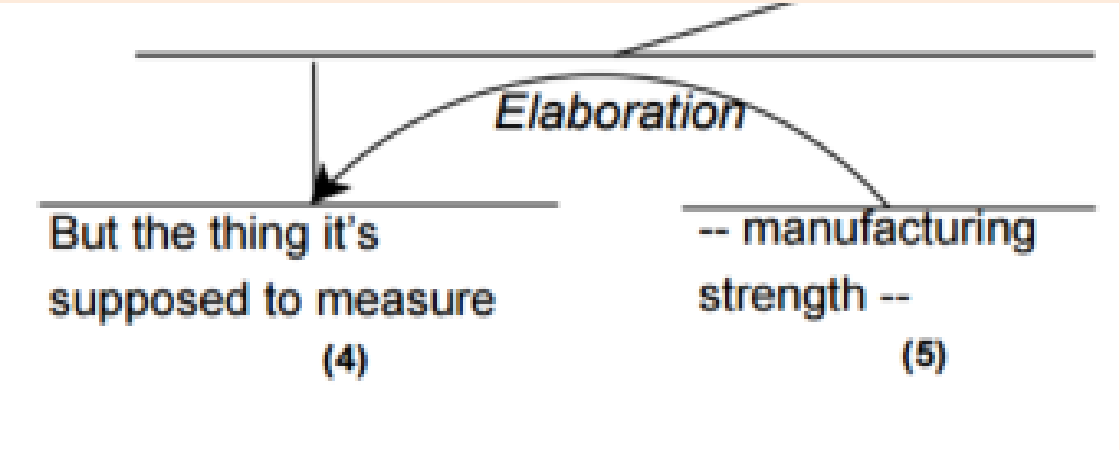
Rhetorical Structure Theory

(Mann and Thomson, 1988)

Describes the structure of a discourse by:

1. Segmenting text into **elementary discourse units** (**EDU**s)
2. Relating spans of text to each other according to a set of **rhetorical relations**
   - Elaboration
   - Attribution
   - Contrast
   - List
   - Background
   - …

Elaboration

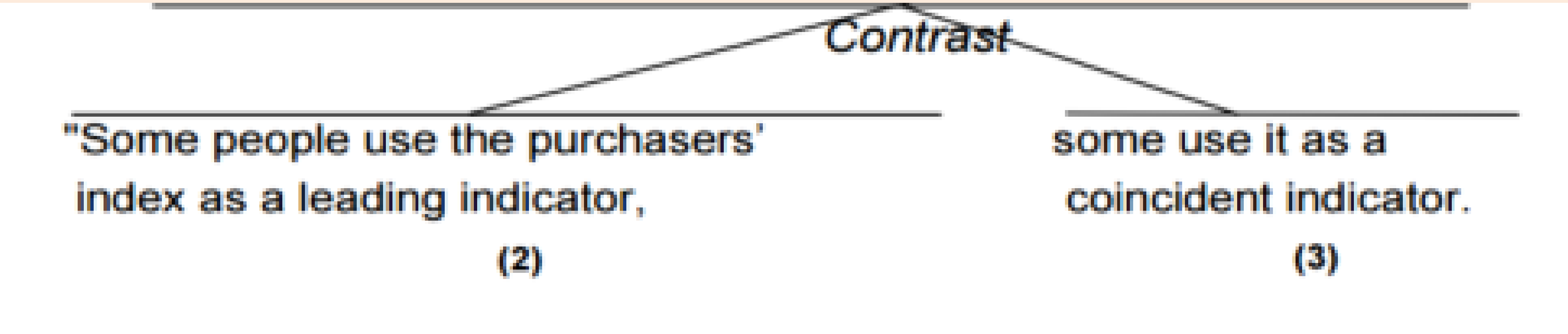


Note that the relation is **asymmetric**:

There is a **nucleus** (4)

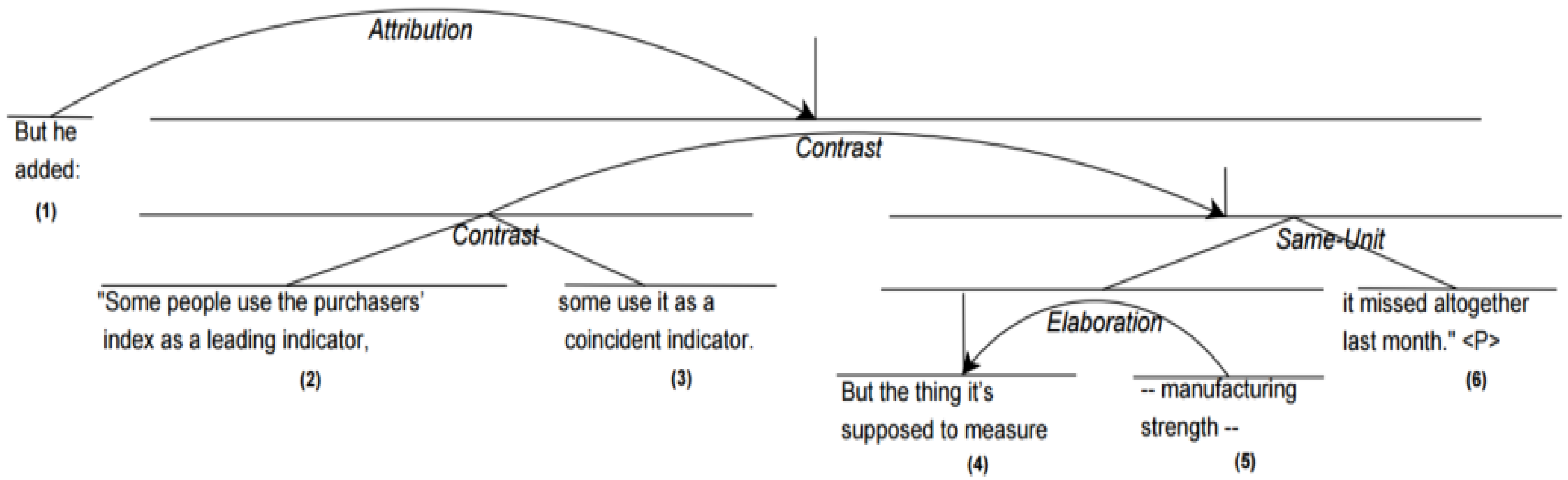
and a **satellite** (5)

Contrast



This particular instance is **symmetric**.

Example of a Full RST Tree



From Joty et al., (2013)

RST Parsing

Usually decomposed into the following steps:

Segmentation of text into EDUs

Recovering the parse structure, with labels

How would you solve this problem?

- How would you decompose the steps?
- What models and algorithms would you use to solve each step?

Applications of RST

People often extract features from RST parse trees to use in downstream applications.

e.g., automatic essay grading

RST is also helpful in automatic summarization.

Marcu (2000) defined heuristics that exploit the asymmetrical structure of RST parse trees to determine the summary content.

Local Coherence Modelling (LCM)

RST builds up a global parse tree, representing the coherence of an entire passage.

Local coherence modelling (Barzilay and Lapata, 2005) emphasizes local cohesive devices that are used to capture coherence between adjacent sentences.

The Life and Death of an Entity

Mentions of entities in a document tend to follow certain patterns:

First mention: often the subject

__Justin Pierre James Trudeau__ MP (born December 25, 1971) is a Canadian politician and the prime minister-designate of Canada. When sworn in, …

Mention clusters – an entity will often appear multiple times within one part of an article, then disappear

... The others are __Ben Mulroney__ (son of Brian Mulroney), Catherine Clark (daughter of Joe Clark), and Trudeau's younger brother, Alexandre. __Ben Mulroney__ was a guest at Trudeau's wedding. …

Ben Mulroney does not appear anywhere else in the article

Entity Grid Model

Make an **entity grid** that plots entity mentions, indicate the syntactic role in which that entity appears:

Table 1
A fragment of the entity grid. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

(Barzilay and Lapata, 2008)

Document Representation

Extract a feature vector representation of a document by taking the relative frequencies of entity mention transitions (i.e., in the style of N-gram models)

Table 3
Example of a feature-vector document representation using all transitions of length two given syntactic categories s, o, x, and –.

	s s	s o	s x	s –	o s	o o	o x	o –	x s	x o	x x	x –	– s	– o	– x	– –
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59
d_2	.02	.01	.01	.02	0	.07	0	.02	.14	.14	.06	.04	.03	.07	0.1	.36
d_3	.02	0	0	.03	.09	0	.09	.06	0	0	0	.05	.03	.07	.17	.39

Evaluations

1. Distinguish original ordering of a document from a version where the ordering of the sentences has been randomly permutated:
   - ~90% accuracy (What is the expected random accuracy?)
2. Evaluate whether one summary is more coherent than another summary
   - ~83% pairwise accuracy
3. Readability assessment: Distinguish *Encyclopedia Britannica* from *Britannica Elementary*.
   - 88.79% accuracy, in conjunction with a pre-existing model for this task.

Extensions

Combining global and local coherence

(Elsner et al., 2007)

Modelling entity relatedness

(Filippova and Strube, 2007)

Other languages than English

(Cheung and Penn, 2010)

Review

Over the past 9 weeks, we have covered computational modelling of:

- Morphology
- Syntax
- Semantics
- Discourse and pragmatics

Computational tools and techniques:

- Machine learning
- Formal language theory

Computational Tasks

Morphological recognition

Is this a well formed word?

Stemming

Cut affixes off to find the stem

- *airliner -> airlin*

Morphological analysis

Lemmatization – remove inflectional morphology and recover lemma (the form you'd look up in a dictionary)

- *foxes -> fox*

Full morphological analysis – recover full structure

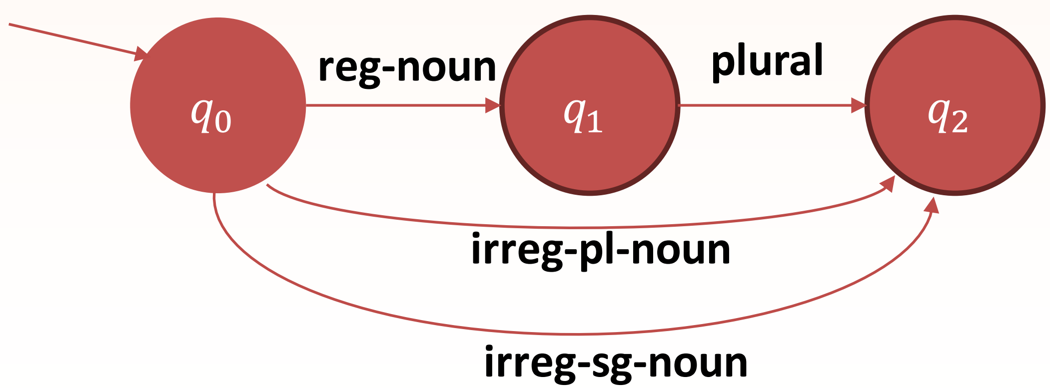
- *foxes -> fox* +N +PL

Definition of FSA

A FSA consists of:

- Q finite set of states
- Σ set of input symbols
- $q_0 \in Q$ starting state
- $\delta : Q \times \Sigma \rightarrow Q$ transition function from current state and input symbol to next state
- $F \subseteq Q$ set of accepting, final states

Identify the above components:



N-grams as Linguistic Knowledge

N-grams can crudely capture some linguistic knowledge and even facts about the world

- e.g., $P(\text{English}|\text{want}) = 0.0011$
  $P(\text{Chinese}|\text{want}) = 0.0065$

  World knowledge: culinary preferences?

  $P(\text{to}|\text{want}) = 0.66$
  $P(\text{eat}|\text{to}) = 0.28$
  $P(\text{food}|\text{to}) = 0$

  Syntax

  $P(\text{I}|\text{<start-of-sentence>}) = 0.25$

  Discourse

MLE Derivation for a Bernoulli

Maximize the log likelihood:

$$\log P(C;\theta) = \log(\theta^{N_1}(1 - \theta)^{N_0})$$
$$= N_1 \log\theta + N_0 \log(1 - \theta)$$

$$\frac{d}{d\theta}\log P(C;\theta) = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0$$

$$\frac{N_1}{\theta} = \frac{N_0}{1 - \theta}$$

Solve this to get:

$$\theta = \frac{N_1}{N_0 + N_1}$$

Or,

$$\theta = \frac{N_1}{N}$$

Supervised vs. Unsupervised Learning

How much information do we give to the machine learning model?

Supervised – model has access to some input data, and their corresponding output data (e.g., a label)

- Learn a function y = f(x), given examples of (x, y) pairs

input

output

model

Unsupervised – model only has the input data

- Given only examples of x, find some interesting patterns in the data

Sequence Labelling

Predict labels for *an entire sequence of inputs*:

? ? ? ? ? ? ? ? ? ? ?

Pierre Vinken , 61 years old , will join the board …

↓

NNP NNP , CD NNS JJ , MD VB DT NN

Pierre Vinken , 61 years old , will join the board …

Must consider:

Current word

Previous context

Forward Algorithm

Trellis of possible state sequences

States	O_1	O_2	$O_{\mathbf{3}}$	$O_{\mathbf{4}}$	$O_{\mathbf{5}}$
VB	$\alpha_{VB}(1)$	$\alpha_{VB}(2)$	$\alpha_{VB}(3)$	$\alpha_{VB}(4)$	$\alpha_{VB}(5)$
NN	$\alpha_{NN}(1)$	$\alpha_{NN}(2)$	$\alpha_{NN}(3)$	$\alpha_{NN}(4)$	$\alpha_{NN}(5)$
DT	$\alpha_{DT}(1)$	$\alpha_{DT}(2)$	$\alpha_{DT}(3)$	$\alpha_{DT}(4)$	$\alpha_{DT}(5)$
JJ	$\alpha_{JJ}(1)$	$\alpha_{JJ}(2)$	$\alpha_{JJ}(3)$	$\alpha_{JJ}(4)$	$\alpha_{JJ}(5)$
CD	$\alpha_{CD}(1)$	$\alpha_{CD}(2)$	$\alpha_{CD}(3)$	$\alpha_{CD}(4)$	$\alpha_{CD}(5)$

Time

$\alpha_i(t)$is $P(\boldsymbol{O}_{1:t}, Q_t = i|\theta)$

- Probability of current tag and words up to now

Discriminative Sequence Model

The parallel to an HMM in the discriminative case: **linear-chain conditional random fields (linear-chain CRFs)** (Lafferty et al., 2001)

$$P(Y|X) = \frac{1}{Z(X)} \exp \sum_{t} \sum_{k} \theta_k f_k(y_t, y_{t-1}, x_t)$$

sum over all features

sum over all time-steps

Z(X) is a normalization constant:

$$Z(X) = \sum_{y} \exp \sum_{t} \sum_{k} \theta_k f_k(y_t, y_{t-1}, x_t)$$

sum over all possible sequences of hidden states

Formal Definition of a CFG

A 4-tuple:

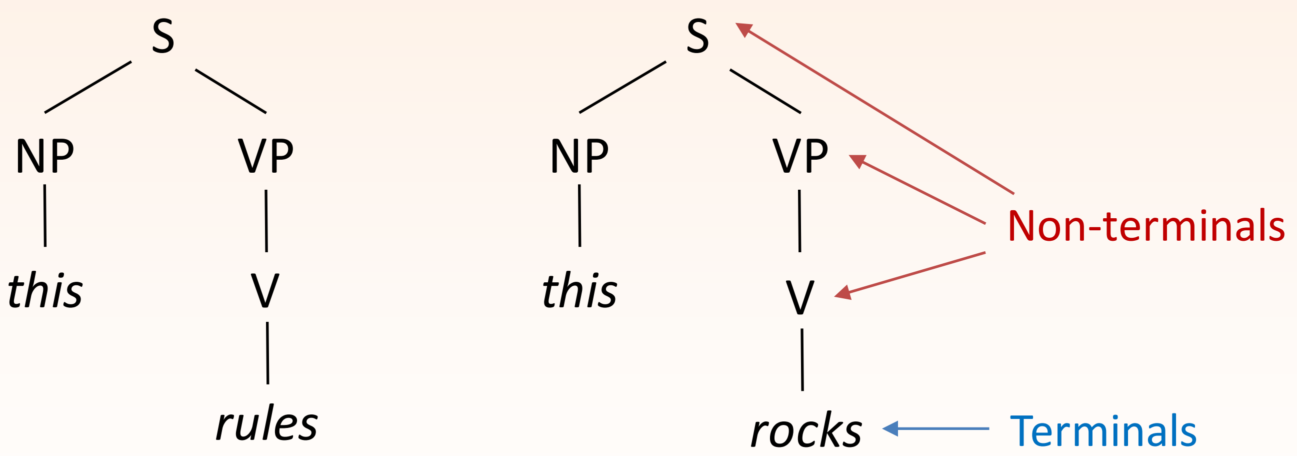
- N set of **non-terminal** symbols
- Σ set of **terminal** symbols
- R set of **rules** or **productions** in the form $A \rightarrow (\Sigma \cup N)^*$, and $A \in N$
- S a designated **start symbol**, $S \in N$

Constituent Tree

Trees (and sentences) generated by the previous rules:

S → NP VP   NP → *this*

VP → V   V → *is* | *rules*| *jumps* | *rocks*



Example of CKY

I_0	$shot_1$	the_2	$elephant_3$	in_4	my_5	$pyjamas_6$
NP, 0.25 N, 0.625 [0:1]	[0:2]	[0:3]	[0:4]	[0:5]	[0:6]	[0:7]
	V, 1.0 [1:2]	[1:3]	[1:4]	[1:5]	[1:6]	[1:7]
		Det, 0.6 [2:3]	NP, ? [2:4]	[2:5]	[2:6]	[2:7]
			NP, 0.1 N, 0.25 [3:4]	[3:5]	[3:6]	[3:7]
				[4:5]	[4:6]	[4:7]
					[5:6]	[5:7]
						[6:7]

New value:

0.6 * 0.25 * Pr(NP → Det N)

Lexical Semantic Relations

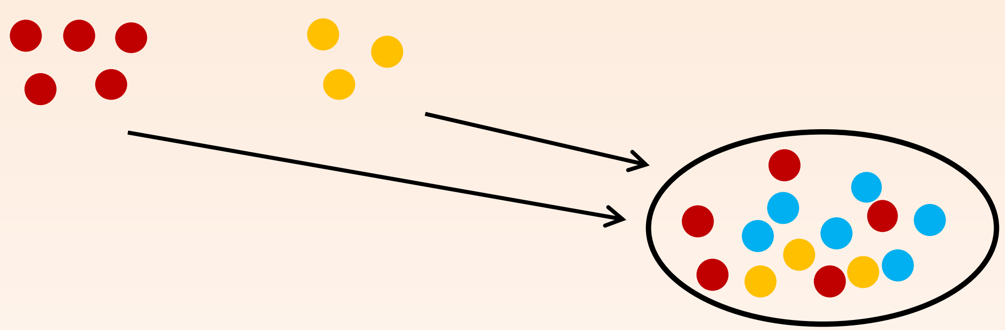
How specifically do terms relate to each other? Here are some ways:

- Hypernymy/hyponymy
- Synonymy
- Antonymy
- Homonymy
- Polysemy
- Metonymy
- Synecdoche
- Holonymy/meronymy

Financial Bank or Riverbank?

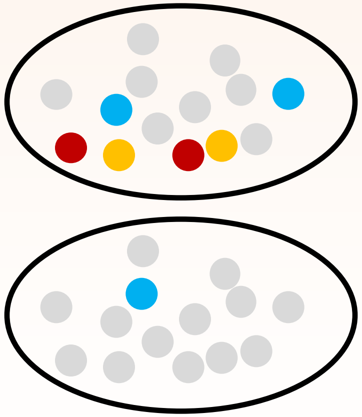
… ***deposit*** *a* ***cheque*** *at the* *bank* *before it* ***closed*** …

- B:



Construct from definitions of **all** senses of context words

- $overlap(bank\#1, B)$
  - 6 overlaps found
- $overlap(bank\#2, B)$
  - 1 overlap found
- Decision: select sense 1.



Yarowsky's Example

Step 1: Disambiguating *plant*

Sense	Training Examples (Keyword in Context)
?	... company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	... zonal distribution of *plant* life . ...
?	... to strain microscopic *plant* life from the ...
?	vinyl chloride monomer *plant* , which is ...
?	and Golgi apparatus of *plant* and animal cells
?	... computer disk drive *plant* located in ...
?	... divide life into *plant* and animal kingdom
?	... close-up studies of *plant* life and natural
?	... Nissan car and truck *plant* in Japan is ...
?	... keep a manufacturing *plant* profitable without
?	... molecules found in *plant* and animal tissue
?	... union responses to *plant* closures . ...
?	... animal rather than *plant* tissues can be
?	... many dangers to *plant* and animal life
?	company manufacturing *plant* is in Orlando ...
?	... growth of aquatic *plant* life in water ...
?	automated manufacturing *plant* in Fremont ,
?	... Animal and *plant* life are delicately
?	discovered at a St. Louis *plant* manufacturing
?	computer manufacturing *plant* and adjacent ...
?	... the proliferation of *plant* and animal life
?	... ...

Term-Context Matrix

Each row is a vector representation of a word

	the	*was*	*and*	*British*	*linguist*
Firth	5	7	12	6	9
figure	276	87	342	56	2
linguist	153	1	42	5	34
1950s	12	32	1	34	0
English	15	34	9	5	21

Context words

Co-occurrence counts

Target words

Frame Semantics (Fillmore, 1976)

Word meanings must relate to **semantic frames**, a schematic description of the stereotypical real-world context in which the word is used.

buy →

Definite Articles

The student took COMP-599:

1. There is an entity who is the student.
2. There is at most one thing being referred to who is a student.
3. The student participates in some predicate.

What is the range of this existential quantifier?

$$\exists x.\, Student(x) \land \forall y.\, (Student(y) \rightarrow y = x)$$
$$\land\, took(x, \mathrm{COMP{-}599})$$

For simplicity, for now, assume took is a predicate, rather than use event variables.

A Course

NP $\lambda u. u(s_1), (\lambda Q. \exists x. Course(x) \land Q(x), 1)$

Det

a

$(\lambda P. \lambda Q. \exists x. P(x) \land Q(x))$

N

course

$(\lambda x. Course(x))$

What is the semantic attachment for NP -> Det N? Use .sem.store to access the store.

Hobb's Algorithm (1978)

A traversal algorithm which requires:

- Constituent parse tree
- Morphological analysis of number and gender

Overall steps:

1. Search the current sentence right-to-left, starting at the pronoun
2. If no antecedent found, search previous sentence(s) left-to-right