

Information Retrieval: between Natural Language, Texts and Meaning (pt. 2)

Roberto Basili

(Università di Roma, Tor Vergata, basili@info.uniroma2.it)

Some slider borrowed from the tutorial «[Natural Language Understanding: Foundations and State-of-the-Art](#)», by [Percy Liang](#) (Stanford University).

Web Mining & Retrieval, a.a. 2020-21

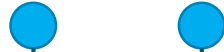
Overview

- Documents in Information Retrieval and Web Applications
- Textual Data, Information and Content
 - Natural Language Processing: introduction to the linguistic background
 - Natural Language and Content
 - NL Syntax
 - NL Semantics
 - Document Representation and IR models
- Summary

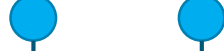


Ambiguity and Linguistic Levels

• Semantics



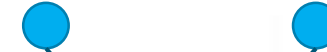
• Syntax



• Morphology



• Phonology



can/can

eat cake with fork

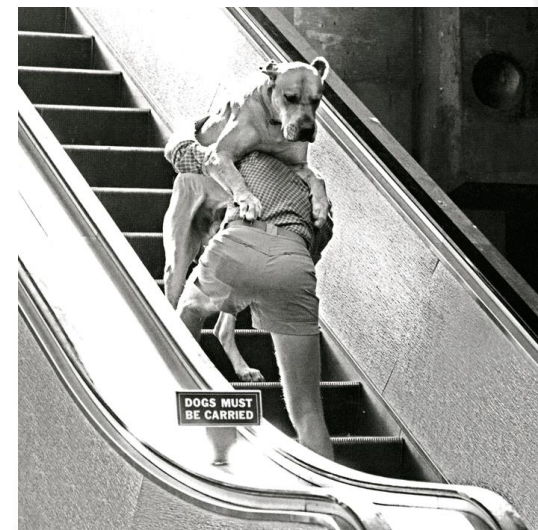
earth observation satellite
Eco's book



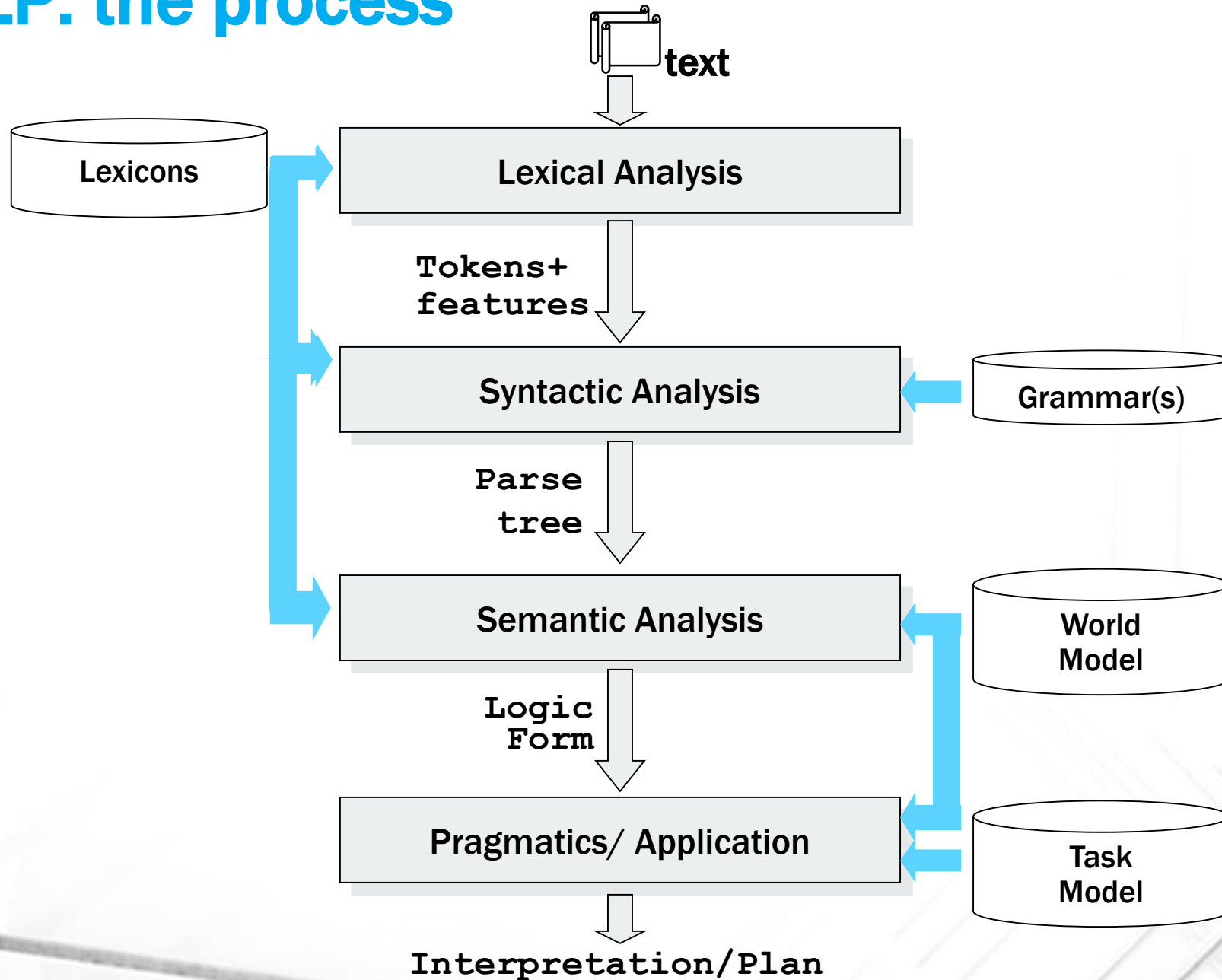
del (pane)
/del (libro)

compro la borsa
in pelle

il timore dei manager



NLP: the process



Syntax

- In linguistics, syntax is *the study of the rules that govern the structure of sentences, and which determine their relative grammaticality.*
- Such rules govern a number of language phenomena as systems for phonology, morphology, syntax as well as discourse

Parse Trees

- The representation of the parsing result is a structure that expresses:
 - The **order of constituent elements** in the sentence
 - The **grammatical type** of constituents
 - The **hierarchical organization of constituents**
- The structures able to express these properties are the **derivation trees** also called **parse trees**

Syntax: Phrase Structure Grammars (Chomsky, 75)

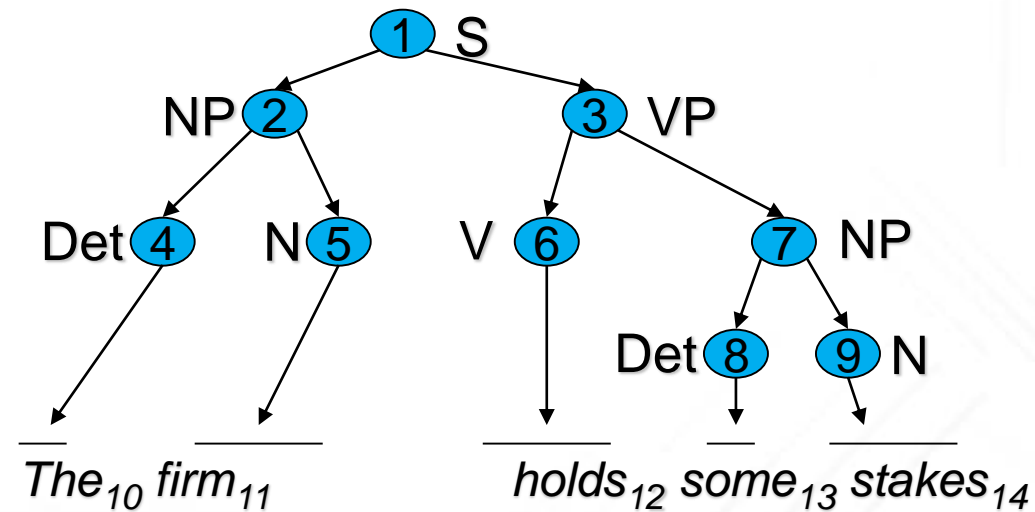
“The firm holds some stakes”

Symbol Vocabulary: $V_n = \{S, NP, VP, Det, N\}$, Axiom: S

Productions: $\{S \rightarrow NP VP, VP \rightarrow V NP, NP \rightarrow Det N\}$

A Derivation is the representation of the cascade of rules used to rewrite S, e.g. :

- $S > NP VP > Det N VP > The N VP > The firm VP > The firm V NP > The firm holds NP > The firm holds Det N > The firm holds some N > The firm holds some stakes$



Grammaticatical Analysis

UK Economy News Headlines - FT.com - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://www.ft.com/world/uk/economy

Google

UK Economy News Headlines - FT.com

Mortgage approvals fall back to January level

Mortgage approvals fell sharply in June, lending yet more weight to the theory of a dip in the UK housing market as the Nationwide index showed UK house prices starting to fall in July - Jul-29

- ▶ Halifax index shows 0.6% fall in house prices
- ▶ In depth: UK house prices
- ▶ House prices rise at slowing rate

Default retirement age to be scrapped

Move delights pressure groups but dismays business organisations, which

Default retirement age to be scrapped

Move delights pressure groups but dismays business organisations, which warn that the measure is being introduced too quickly - Jul-29


Global Insight: Cameron needs to be more subtle

David Cameron has led the largest official delegation to India since its independence from Britain 63 years ago. By doing so, he has tested Britain's place in the world, and how far it has travelled since 1947 - Jul-29

Gilts lose lustre for overseas investors

Flight from eurozone risk to UK government bonds is moderating - Jul-29

UK government spending In depth



Westminster blog
With Alex Barker and Jim Pickard

SEARCH

Enter keywords Go

Regional Business Controller
Consumer Products

UK Business Development Manager -
Building Services Projects
Mechanical & Electrical Engineering

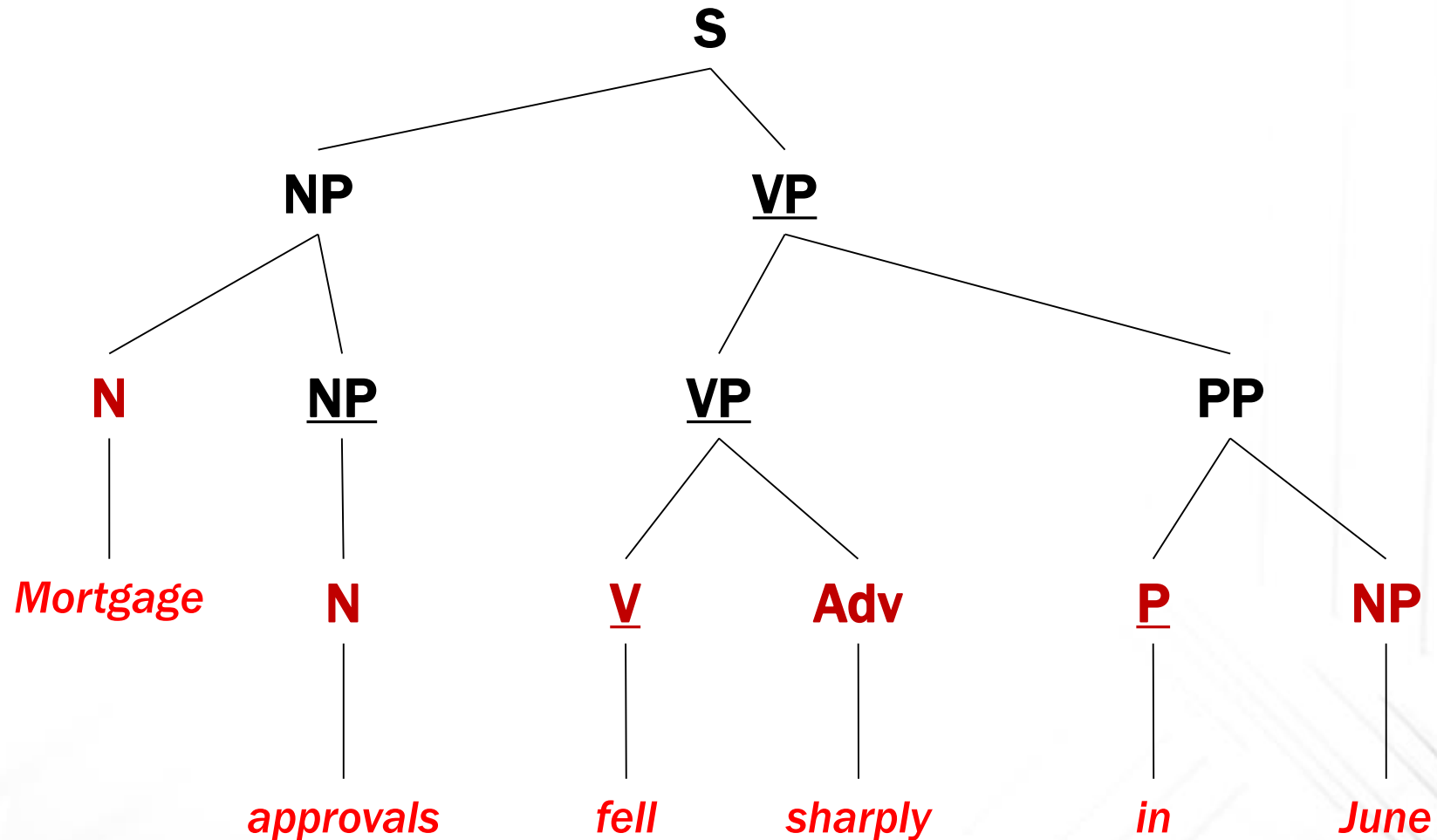
Deputy Director of Finance
London Ambulance Service

RECRUITERS

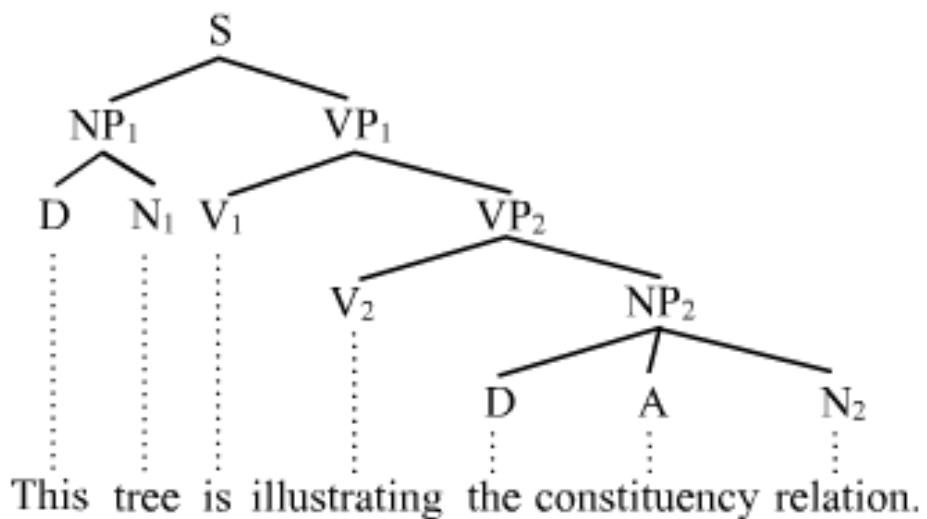
Italiano (Italia)

TexFlame

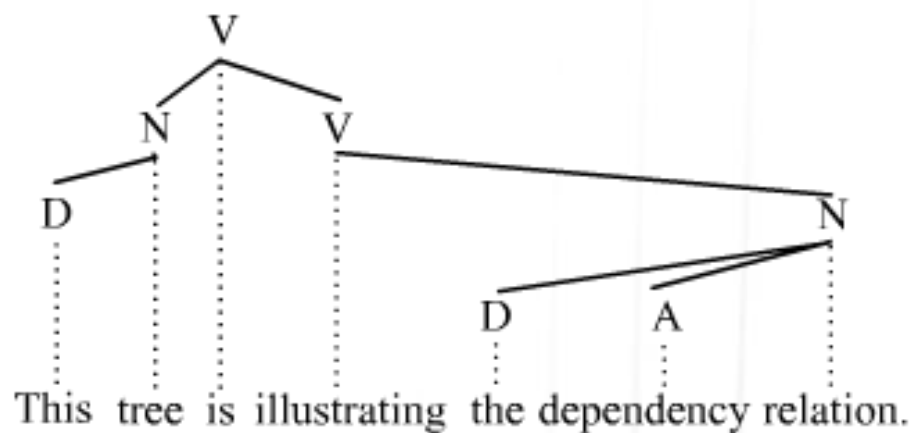
Constituent-based Parsing (with marked Heads)



Constituency-relation vs. Dependency

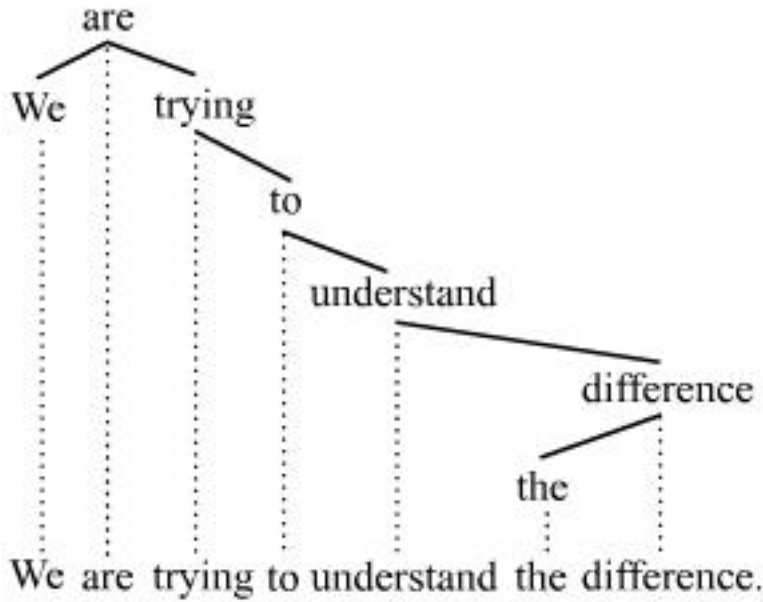


Constituency relation (PSG)

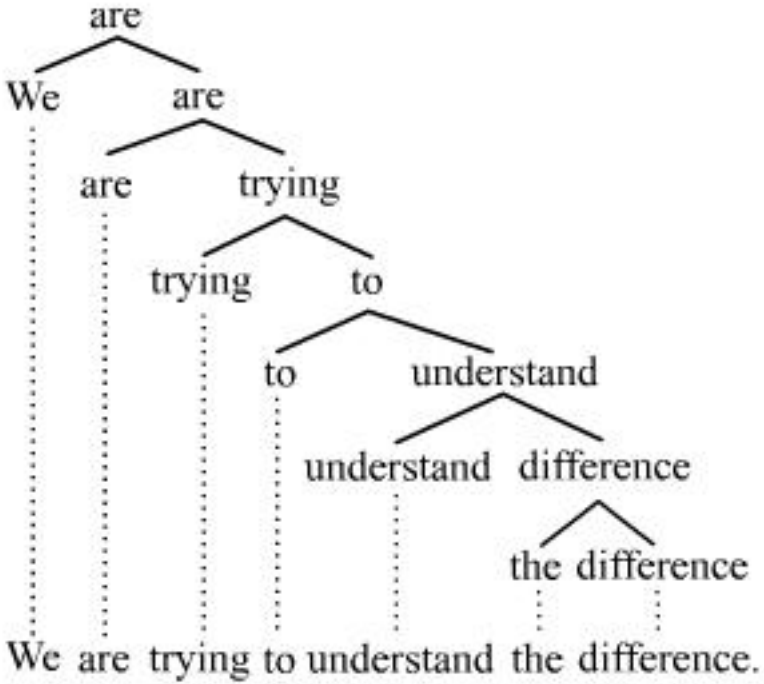


Dependency relation

Constituency vs. Dependency



Dependency



Constituency (BPS)

Dependency Structures

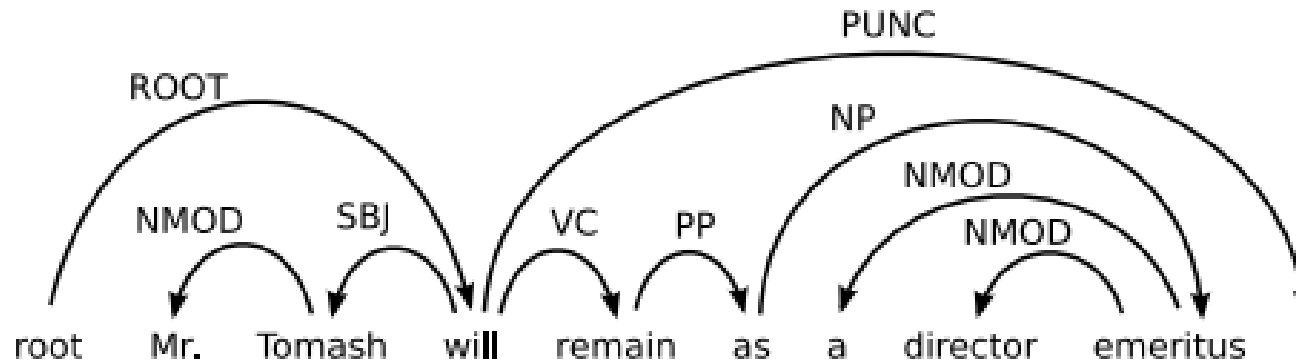


Figure 1: A projective dependency graph.

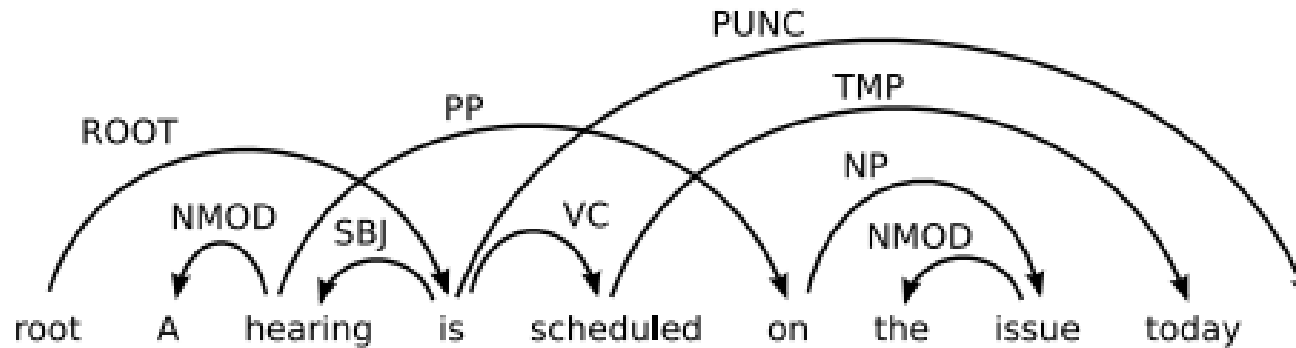
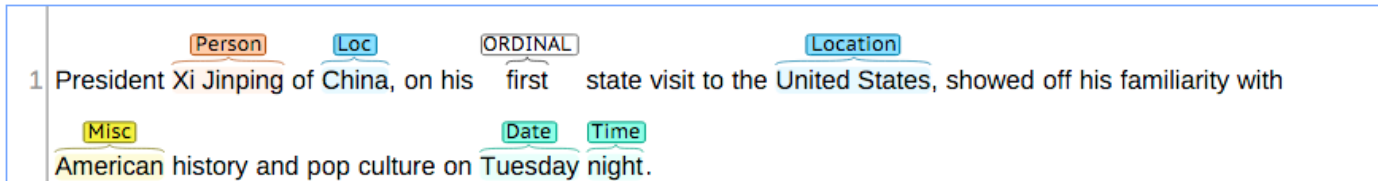


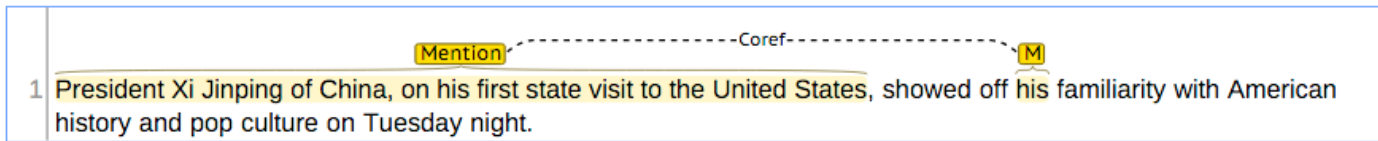
Figure 2: Non-projective dependency graph.

Dependency Parsing

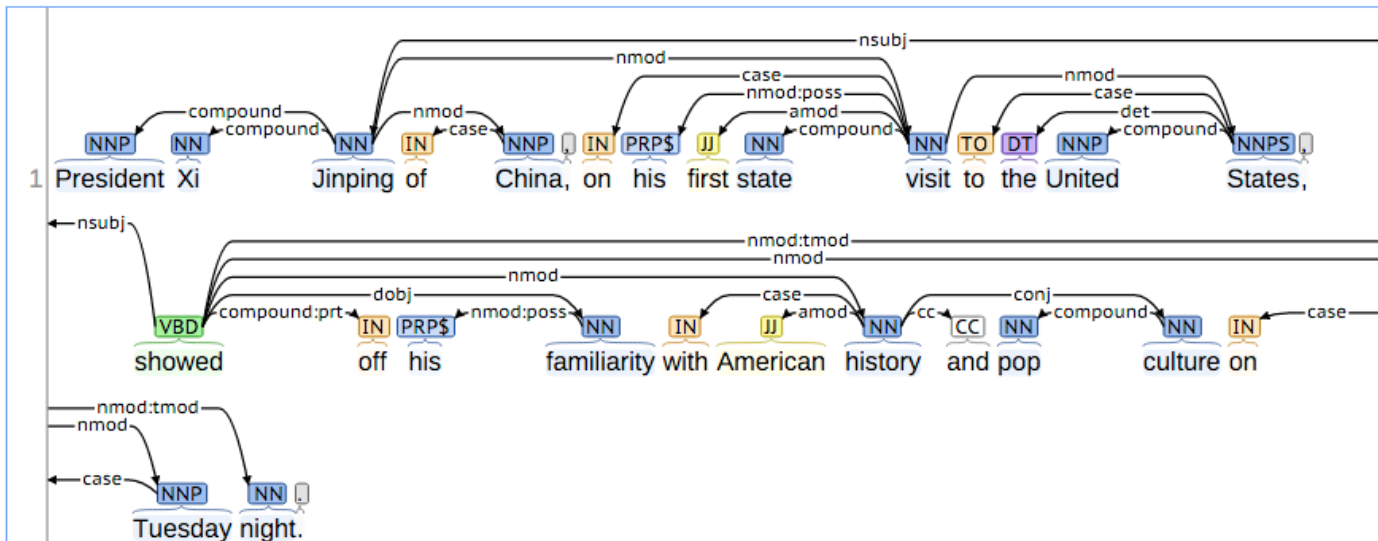
Named Entity Recognition:

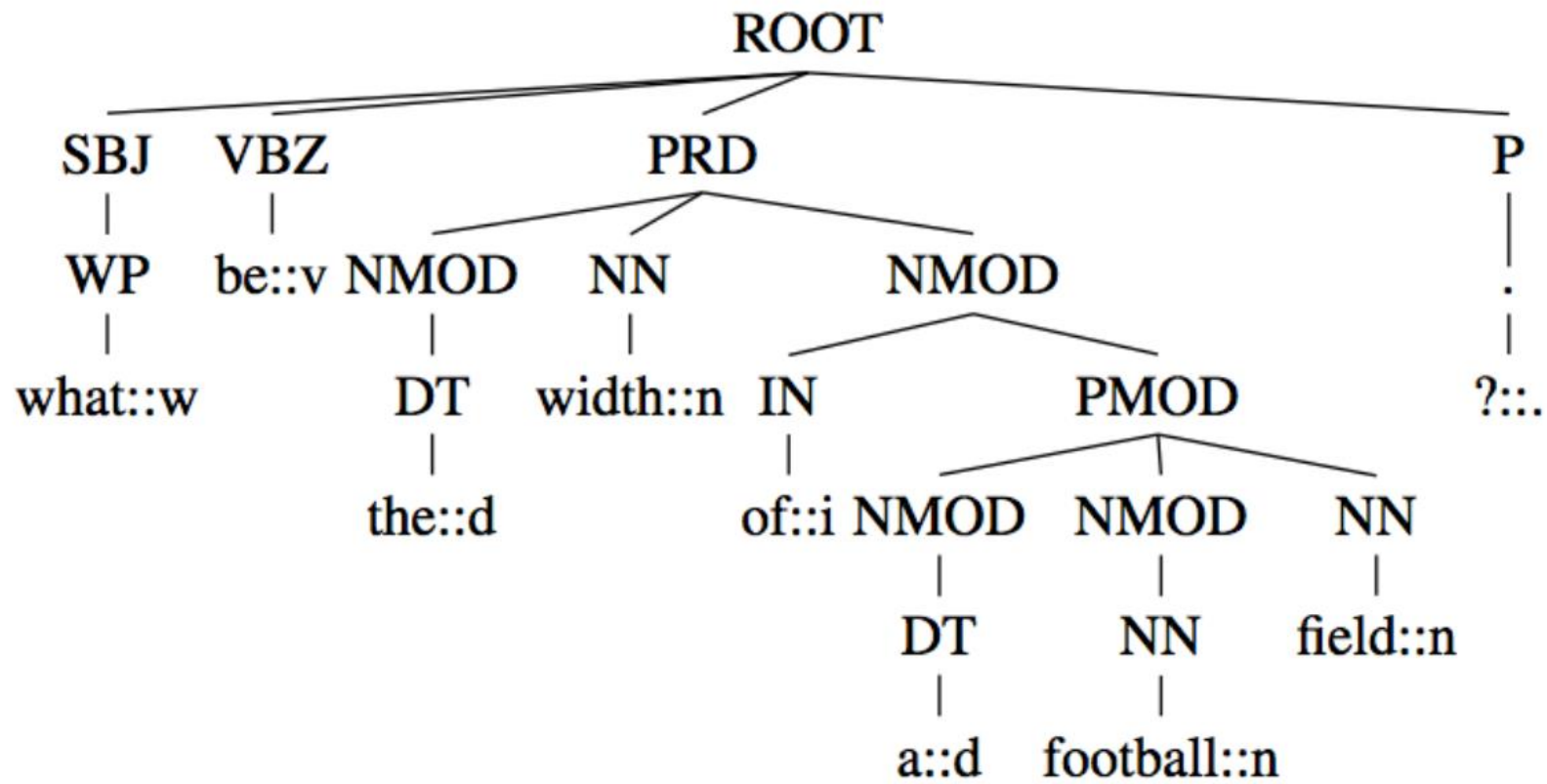


Coreference:



Basic Dependencies:

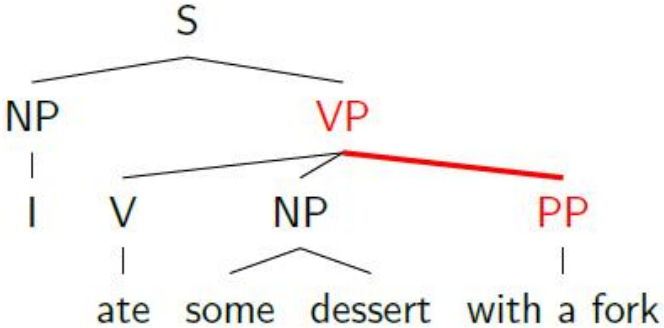
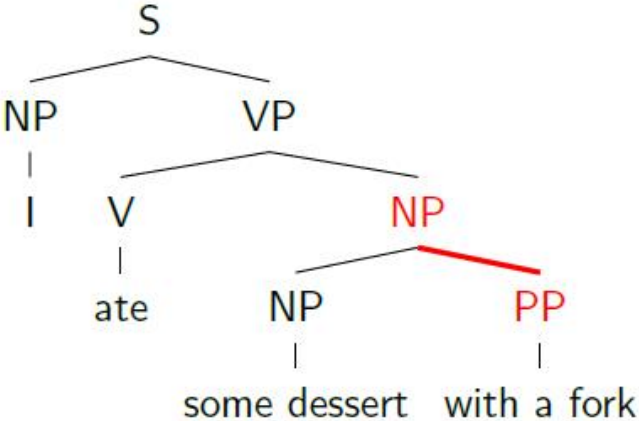




Grammatical Relation Centered Tree
(GRCT)

Grammars & Ambiguity

I ate some dessert with a fork.



Parsing & Ambiguity

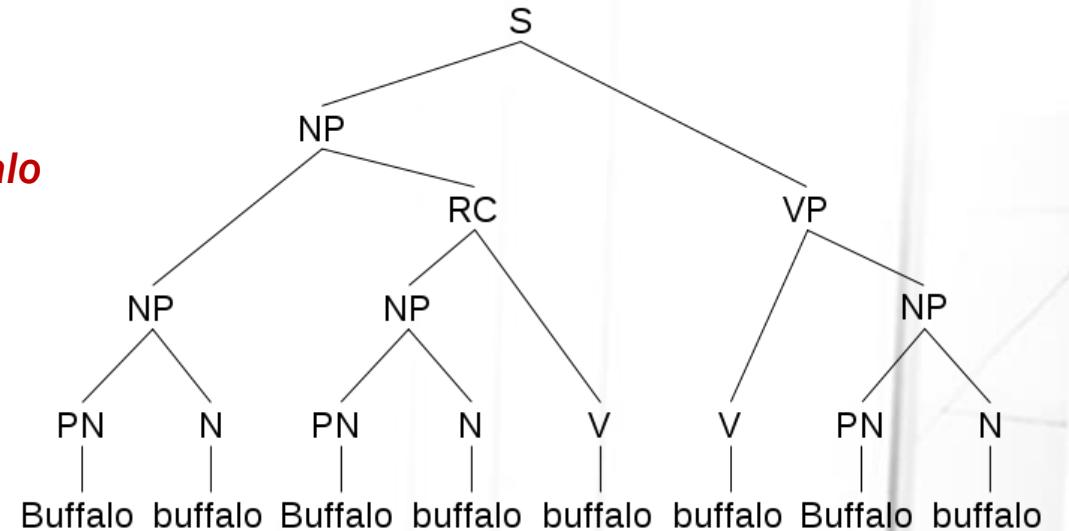
- The parser search space is huge as for the effect of several forms of ambiguity that interacts in a combinatorial way
 - e.g. *La vecchia porta la sbarra*,
 - or *Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo*
- Notice the strong relationship with semantics
 - Most of the ambiguities cannot be solved at the sole syntactic level
 - Lexical information (e.g. word senses) are crucial:



• *To operate in a market* viz. *To operate a body part*



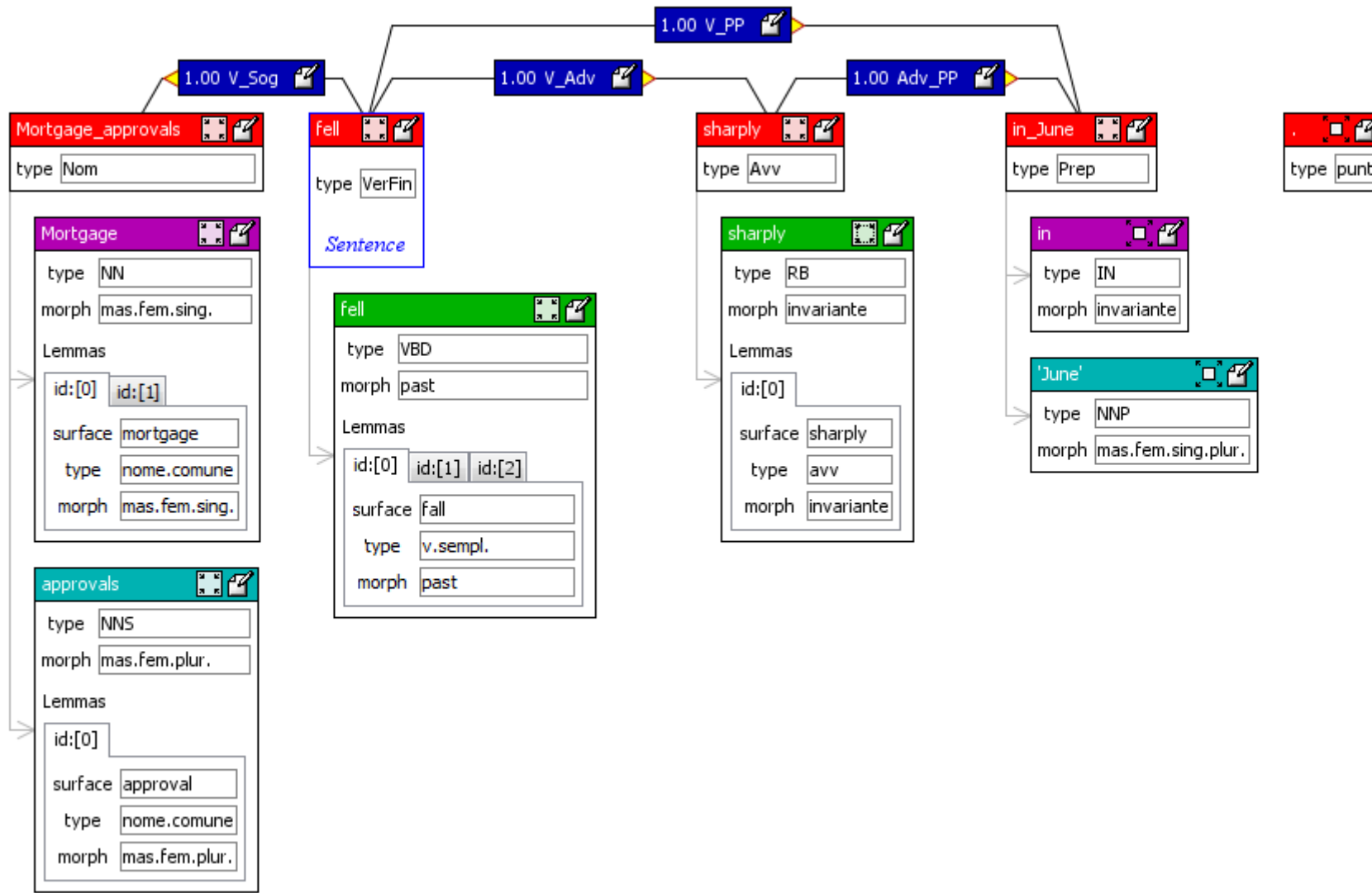
• *Operare in un mercato* ≠ *Operare un paziente*



Bison from Buffalo, New York who are intimidated by other bison in their community also happen to intimidate other bison in their community



**(A(SHIP SHIPPING)SHIP)
SHIPPING(SHIPPING SHIPS))**



FT (July, 29): *Mortgage approvals fell sharply in June.*

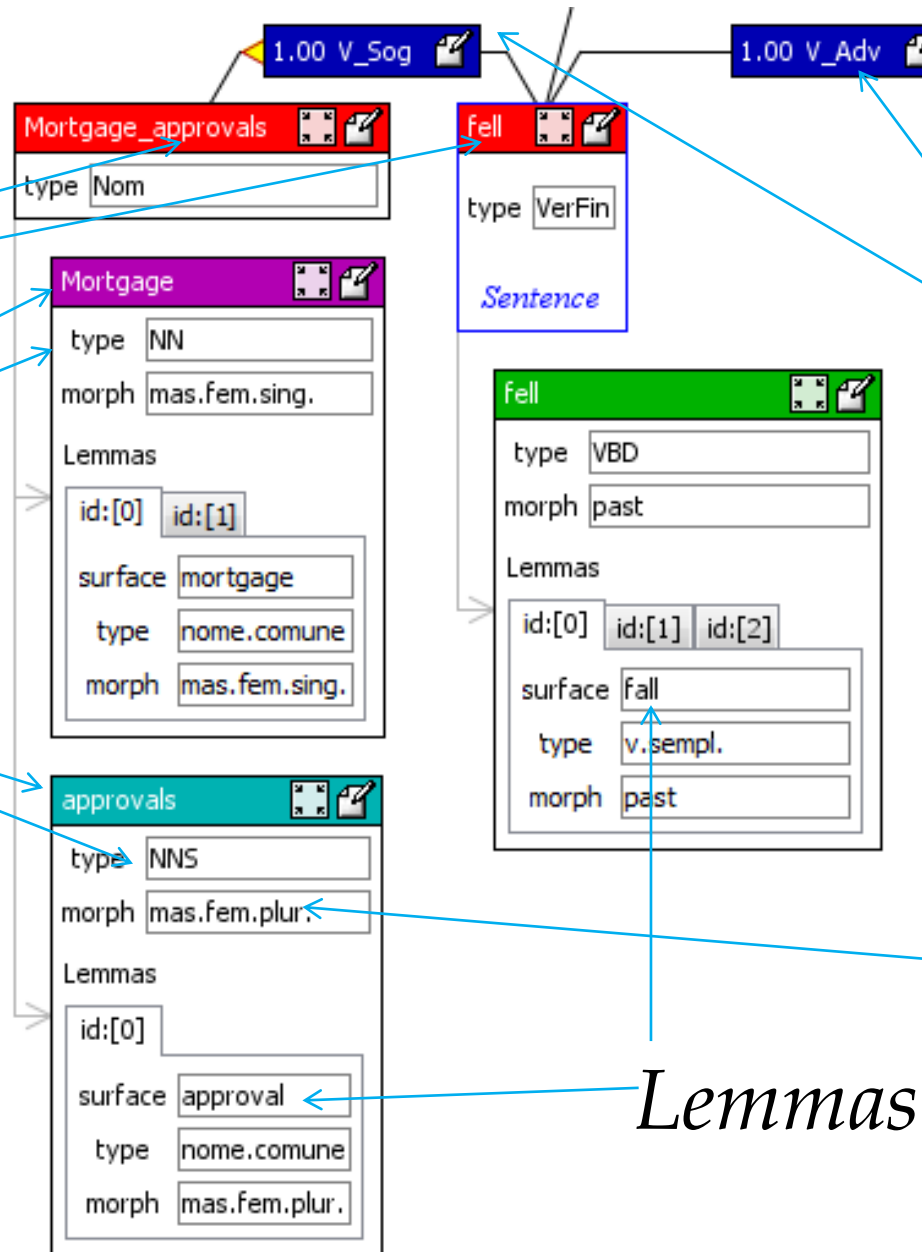
*Tokens and
POS tags*

Chunks

*Grammatical
Relations*

*Morphological
Features*

Lemmas



FT (July, 29): Mortgage approvals fell sharply in June.

Semantics

- What is the meaning of the sentence

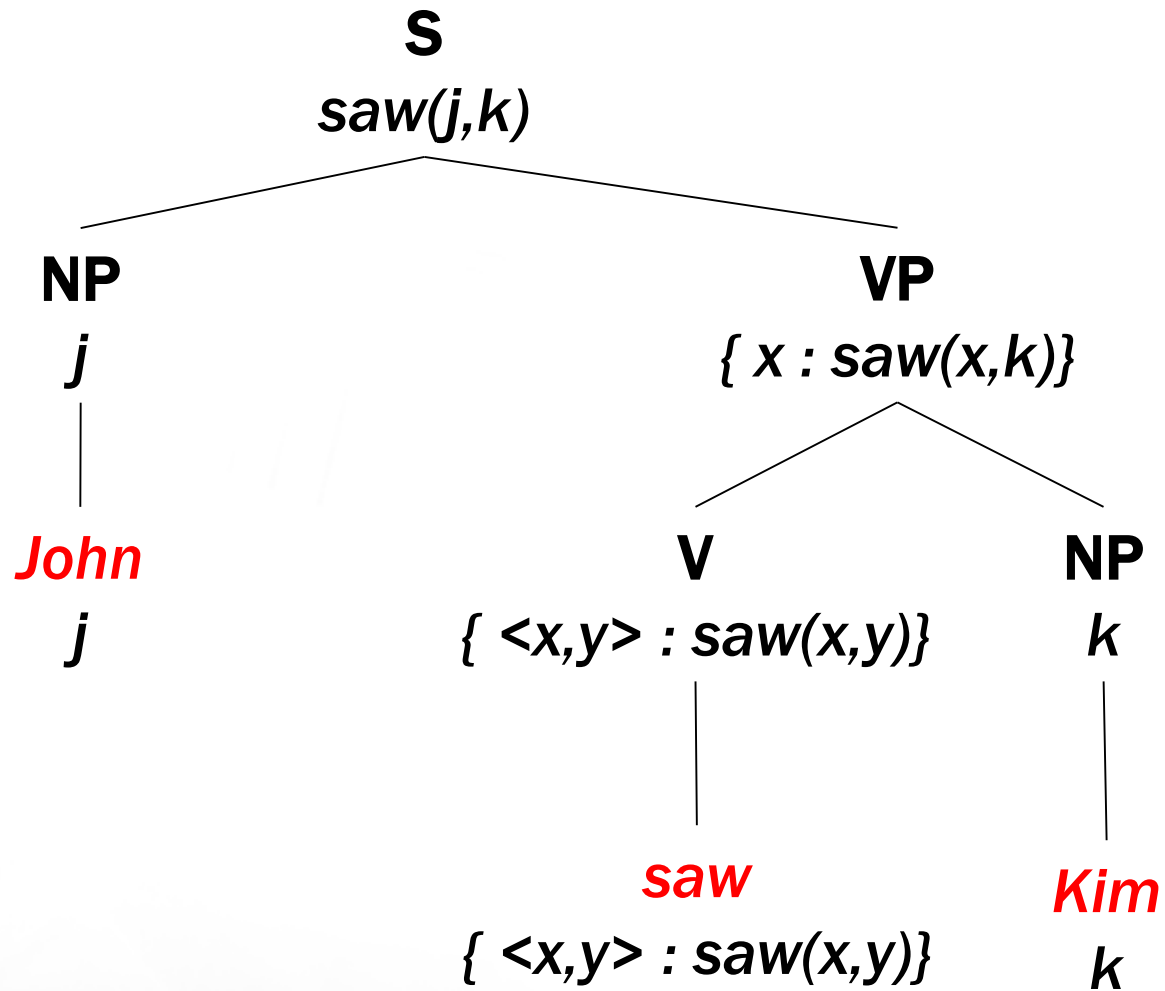
John saw Kim?

- Desirable Properties:

- It should be derivable as a function of the individual constituents, i.e. the meanings of constituents such as Kim, John and see
- Independent from syntactic phenomena, e.g. *Kim was seen by John* is a paraphrasis
- It must be directly used to trigger some inferences:
 - *Who was seen by John?* Kim!
 - *John saw Kim. He started running to her.*



A Truth conditional semantics



John saw Kim

Syntax and Semantics in textual data

- Compositionality

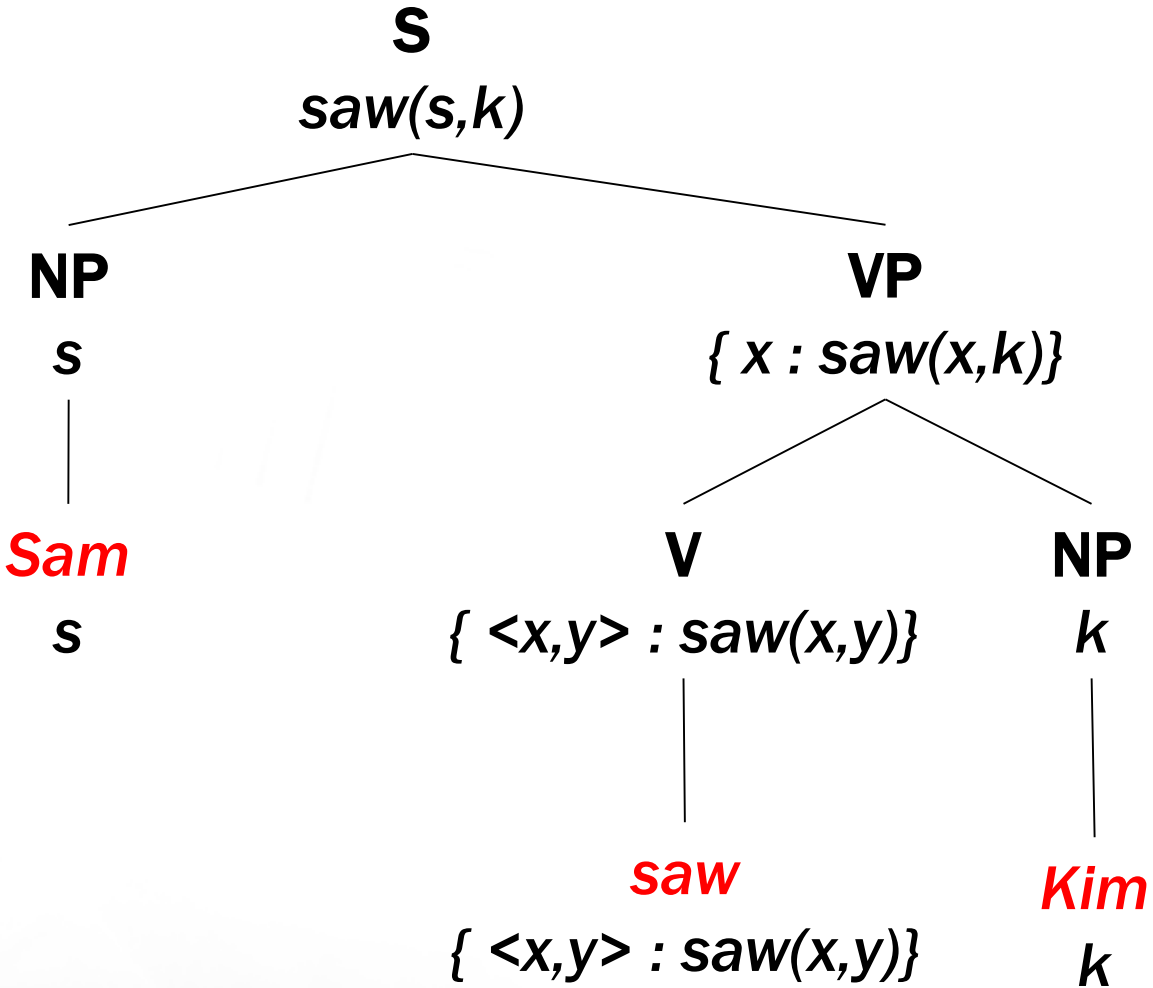
The meaning of a complex expression is solely determined by the meanings of its constituent expressions and the rules used to combine them.

- *"I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages are languages in this sense. Similarly, the set of "sentences" of some formalized system of mathematics can be considered a language"*
Chomsky 1957

A truth-conditional program for NL semantics

- To define a **representation** for the semantics of sentences in natural languages
 - Forme logiche con quantificatori
 - Forme relazionali (ground, i data record delle Basi di dati)
 - Vettori del modello bag-of-words (dei documenti) in stile Rocchio
- To determine **a procedure** for (automatically) generating such a (selected) representation
- To (formally) **support the different inferences** based on the representation that are harmonic with the ones carried out by speakers and hearers of the language
 - *Dimostrare i teoremi*
 - *Rispondere a query SQL*
 - *CI fanno addestrare i classificatori per categorizzare i testi*

A Truth conditional semantics



Sam saw Kim

Towards Lambda-calculus

- *Giuseppe runs produrrebbe: run(Giuseppe)*
- ² *Every student writes programs*

$$\forall x \text{ student}(x) \Rightarrow (\exists p)(\text{program}(p) \& \text{write}(x,p))$$

- *Reflection:*
- VP map towards *predicates (predicative symbols)*
- Proper Nouns map into (ground) atomic symbols
- Quantification require more complex structures
- Logical forms corresponding to VP (VP') are functions from entities to propositions

Functions and lambda-calculus

- $f(x) = x+1$
- A better abstraction about f can be obtained as follows: $\lambda x.x+1$
 - $(\lambda x.x+1)(3)$ $((\lambda x.(x+1))(3))$ corresponds to $3+1$
- Main consequences
 - There is no need of names for functions
 - Operations Ω needed to compute a function f are explicit
- β -reduction: $(\lambda x.\Omega)a \rightarrow [\Omega]\{x = a\}$
- while,
 - $(\lambda x.\lambda y.\Omega)(a)(b) = \lambda y.\Omega\{x=a\}(b) = [\Omega]\{x = a, y = b\}$

λ -Calculus: Syntax

If ϕ is a formula and v a variable then $\lambda v.\phi$ is a predicate. In general, if ψ is an n -ary predicate and v is a variable, then $\lambda v.\psi$ is an $n + 1$ -ary predicate.

- $\lambda x.run(x)$
- $\lambda x.see(x, g)$
- $\lambda x.see(m, x)$
- $\lambda y.\lambda x.see(x, y)$

λ -Calculus: Semantics

If ϕ is a formula and v a variable then the semantics of $\lambda v.\phi$ is the characteristic (membership) function of the set of entities that **satisfy** ϕ (i.e. they make it true).

- $\lambda x.run(x)$
- $\lambda x.see(x, g)$
- $\lambda x.see(m, x)$
- $\lambda y.\lambda x.see(x, y)$

β -reduction and Compositional Semantics

The following expressions are equivalent:

$(\lambda x.run(x)) (g)$	$run(g)$
$(\lambda x.see(x, g))(m)$	$see(m, g)$
$(\lambda x.see(m, x))(g)$	$see(m, g)$

In this framework, the computation of the (compositional) semantics of a sentence is mapped into a recursive application of functions (i.e. lambda-expressions) associated to the grammatical symbols.

β -reduction

The *beta*-reduction $(\lambda x.\Omega)a$ is carried out by substituting contemporarily **all** the (free) occurrences of the variable x in Ω with the expression a .

Operation	Λ -Expression	Result
β -reduction:	$(\lambda x.\Omega)a$	$[\Omega]\{x = a\}$
	$(\lambda x.\lambda y.\Omega)(a)(b)$	$\lambda y.\Omega\{x = a\}(b) = [\Omega]\{x = a, y = b\}$

β -reduction and Compositional Semantics

- *Giuseppe runs: run(giuseppe)*
- $S \rightarrow NP VP$
- Sem Rule1 (*Intransitive verbs*):
IF The Logical Form (LF) of NP is NP' and the LF of VP is VP' :
THEN the LF S' corresponds to VP'(NP')
- Consequences:
runs: $\lambda x.run(x)$
Giuseppe: giuseppe
- $S' = VP'(NP') = (\lambda x.run(x))(giuseppe) = run(giuseppe)$

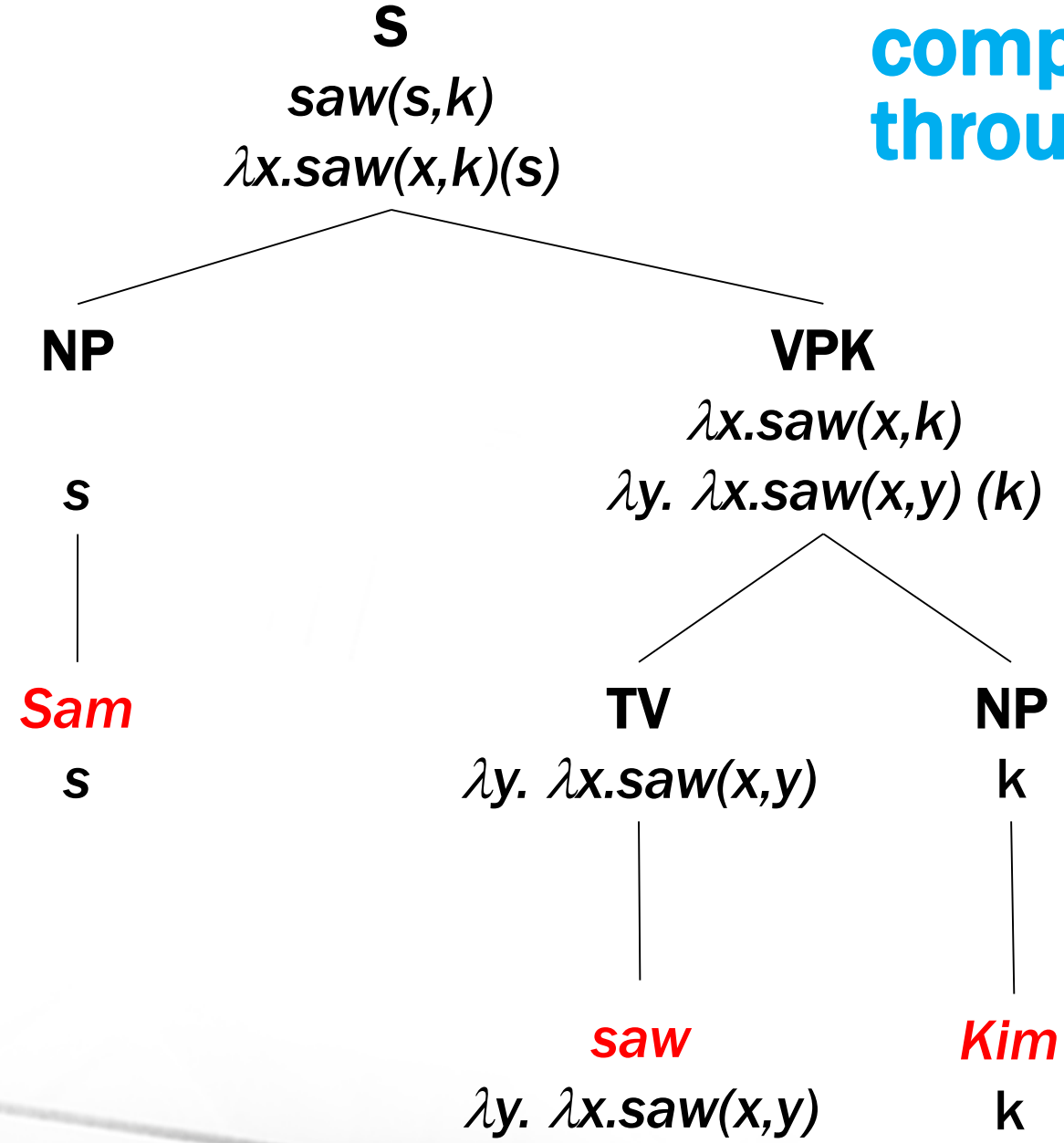
**Example:
transitive verbs**

β -reduction and Compositional Semantics (2)

Example: transitive verbs

- *Giuseppe knows Prolog*: $know(giuseppe, prolog)$
- $VP \rightarrow V NP$
- Sem Rule2 (*transitive verbs*):
IF the LF of NP is NP' and the LF of V is V' :
THEN the LF of VP' corresponds to $V'(NP')$
- Consequences (in the semantic modelling V' of a verb phrase):
 $knows: \lambda x.\lambda y.know(y, x)$
- $S' = VP'(NP'_0) =$
 $= V'(NP'_1)(NP'_0) = (\lambda x.\lambda y.know(y, x))(prolog)(giuseppe) =$
 $= know(giuseppe, prolog)$

NL Interpretation as compositional processing through *lambda expressions*



Sam watched Kim

Sam was screeing Kim

Kim was seen by Sam

John's son watched Kim

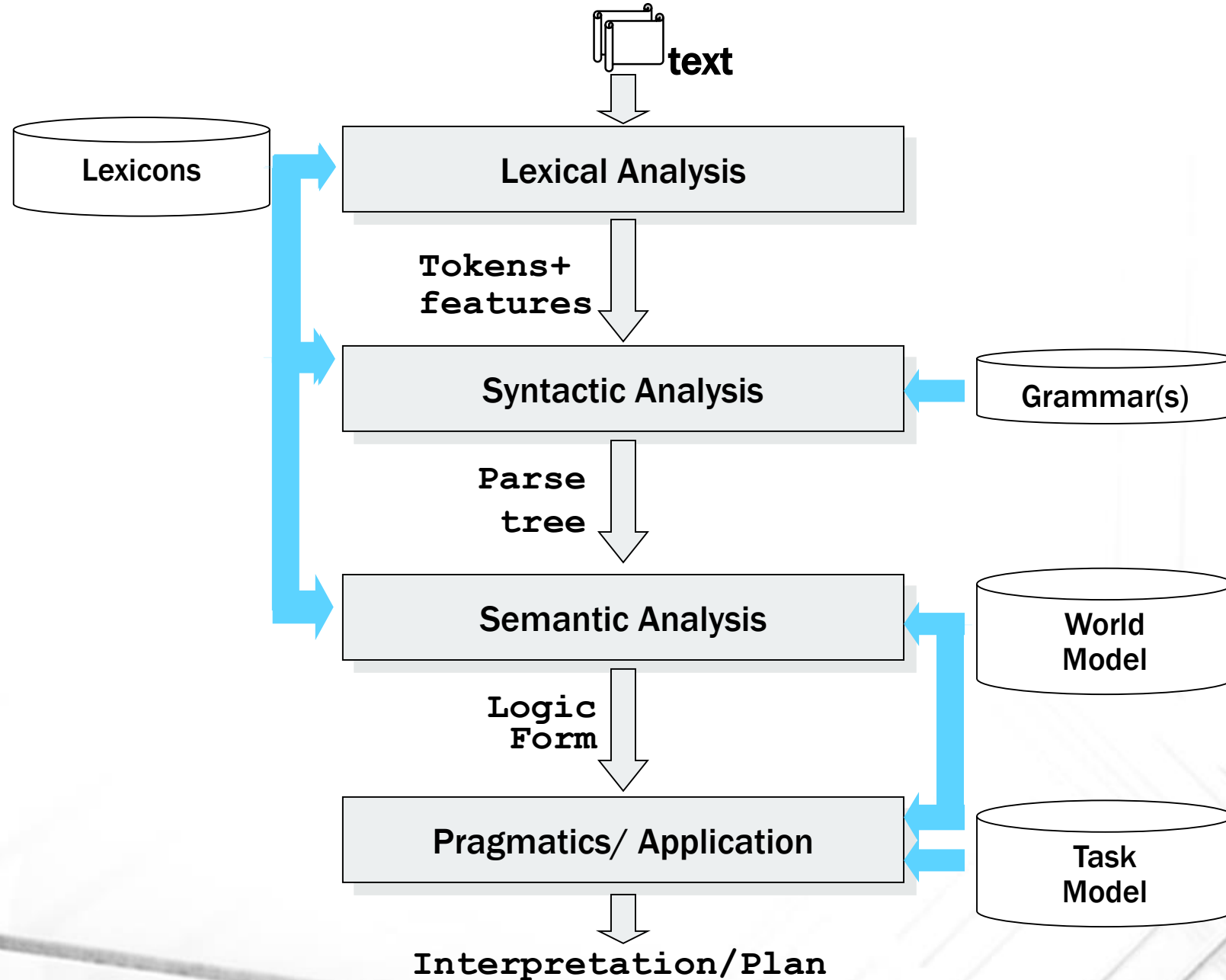
Sam saw Jane's daughter

saw (s , k)

?-saw (X , k) .

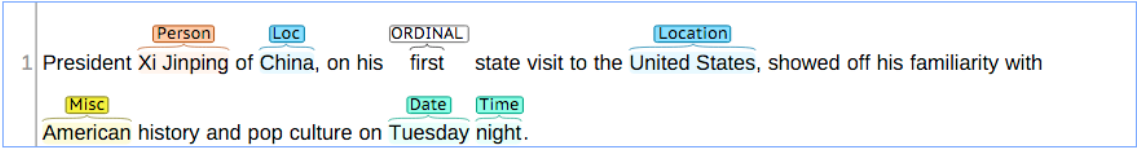
X=s

NLP: the standard processing chain

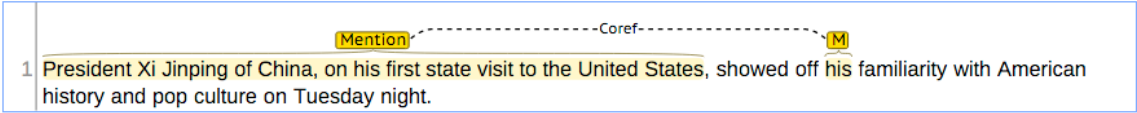


Beyond Parsing: Named Entity Recognition & Coreference

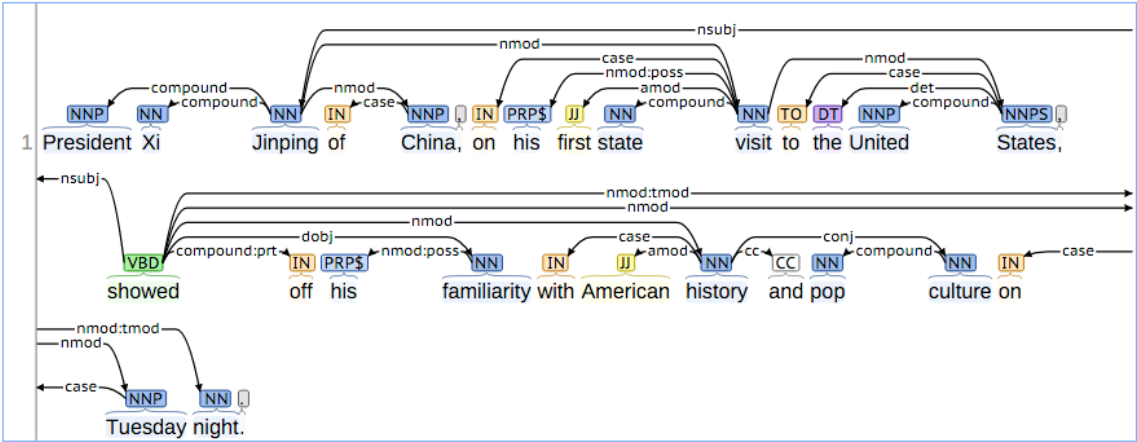
Named Entity Recognition:



Coreference:



Basic Dependencies:



Three Linguistic Perspectives on Meaning

- **Lexical Semantics**
 - The meanings of individual words
- **Formal Semantics (or Compositional Semantics or Sentential Semantics)**
 - How those meanings combine to make meanings for individual sentences or utterances
- **Discourse or Pragmatics**
 - How those meanings combine with each other and with other facts about various kinds of context to make meanings for a text or discourse
 - Dialog or Conversation is often lumped together with Discourse

Lexical Semantic: Relationships between word meanings

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernymy
- Hyponymy
- Meronymy

Homonymy

- Homonymy:
 - Lexemes that share a form
 - Phonological, orthographic or both
 - But have unrelated, distinct meanings
 - Clear example:
 - *Bat* (wooden stick-like thing) vs
 - *Bat* (flying scary mammal thing)
 - Or *bank* (financial institution) versus *bank* (riverside)
 - Can be also homophones, homographs, or both:
 - Homophones:
 - *Write* and *right*
 - *Piece* and *peace*

Polysemy

- *The bank is constructed from red brick*
- *I withdrew the money from the bank*
- Are those the same sense?
- Or consider the following WSJ example
 - *While some banks furnish sperm only to married women, others are less restrictive*
- Which sense of *bank* is this?
 - Is it distinct from (homonymous with) the *river bank* sense?
 - How about the *savings bank* sense?

Synonyms

- Word that have the same meaning in some or all contexts.
 - *filbert / hazelnut*
 - *couch / sofa*
 - *big / large*
 - *automobile / car*
 - *vomit / throw up*
 - *Water / H₂O*
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
 - If so they have the same propositional meaning

Synonyms

- But there are few (or no) examples of perfect synonymy.
 - Why should that be?
 - Even if many aspects of meaning are identical still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
 - *Water and H2O*
 - I would not say
 - *I like fresh **H2O** after the tennis*

Some terminology

- Lemmas and wordforms
 - A lexeme is an abstract pairing of meaning and form
 - A lemma or citation form is the grammatical form that is used to represent a lexeme.
 - *Carpet* is the lemma for *carpets*, *Dormir* is the lemma for *duermes*.
 - Specific surface forms *carpets*, *sung*, *duermes* are called wordforms
- The lemma *bank* has two senses:
 - *Instead, a bank can hold the investments in a custodial account in the client's name*
 - *But as agriculture burgeons on the east bank, the river will shrink even more.*
- **A sense is a discrete representation of one aspect of the meaning of a word**

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
 - *How big is that plane?*
 - *Would I be flying on a large or small plane?*
- How about here:
 - *Miss Nelson, for instance, became a kind of big sister to Benjamin.*
 - *?Miss Nelson, for instance, became a kind of large sister to Benjamin.*
- Why?
 - *big* has a sense that means *being older, or grown up*
 - *large* lacks this sense

II. WordNet (Miller, 1991)

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
 - Versions for other languages are under development

Category	Unique Forms
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

WordNet

- Home page: <http://wordnetweb.princeton.edu/perl/webwn>

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- [S: \(n\) meaning](#), [significance](#), [signification](#), [import](#) (the message that is intended or expressed or signified) "*what is the meaning of this sentence*"; "*the significance of a red traffic light*"; "*the signification of Chinese characters*"; "*the import of his announcement was ambiguous*"
- [S: \(n\) meaning](#), [substance](#) (the idea that is intended) "*What is the meaning of this proverb?*"

Verb

- [S: \(v\) mean](#), [intend](#) (mean or intend to express or convey) "*You never understand what I mean!*"; "*what do his words intend?*"

WordNet

- Home page: <http://wordnetweb.princeton.edu/perl/webwn>

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- **S: (n) meaning, [significance](#), [signification](#), [import](#)** (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
 - [direct hyponym](#) / [full hyponym](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [derivationally related form](#)
- **S: (n) meaning, [substance](#)** (the idea that is intended) *"What is the meaning of this proverb?"*

Wordnet: hyponyms of the word sense meaning₁

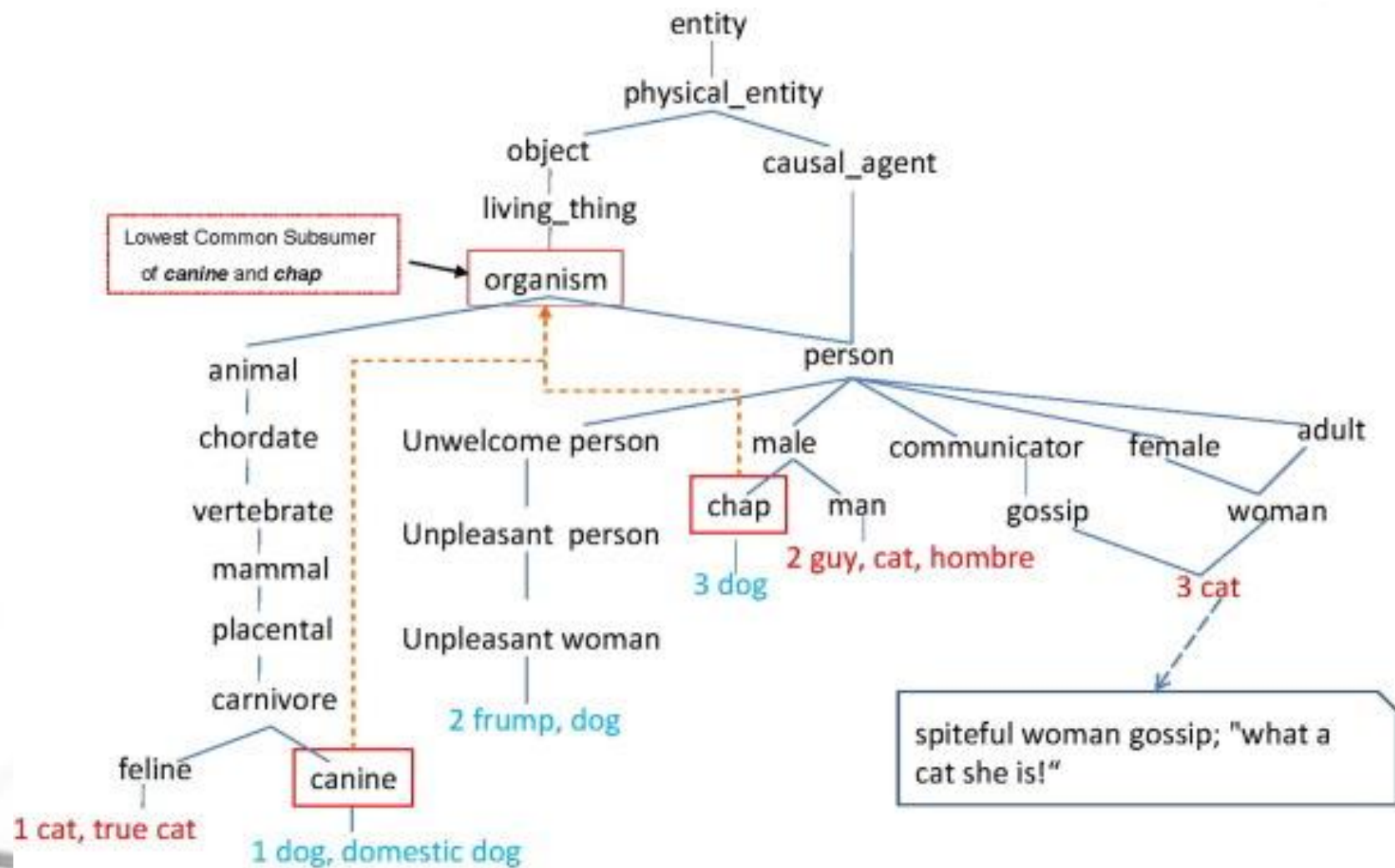
- S: (n) meaning, significance, signification, import (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
 - direct hyponym / full hyponym
 - S: (n) lexical meaning (the meaning of a content word that depends on the nonlinguistic concepts it is used to express)
 - S: (n) grammatical meaning (the meaning of a word that depends on its role in a sentence; varies with inflectional form)
 - S: (n) symbolization, symbolisation (the use of symbols to convey meaning)
 - S: (n) sense, signified (the meaning of a word or expression; the way in which a word or expression or situation can be interpreted) *"the dictionary gave several senses for the word"; "in the best sense charity is really a duty"; "the signifier is linked to the signified"*
 - S: (n) intension, connotation (what you must know in order to determine the reference of an expression)
 - S: (n) referent (something referred to; the object of a reference)
 - S: (n) effect, essence, burden, core, gist (the central meaning or theme of a speech or literary work)
 - S: (n) intent, purport, spirit (the intended meaning of a communication)
 - S: (n) moral, lesson (the significance of a story or event) *"the moral of the story is to love thy neighbor"*

Wordnet: hyperonyms of the word sense *meaning*₁

Noun

- S: (n) meaning, significance, signification, import (the message that is intended or expressed or signified) "*what is the meaning of this sentence*"; "*the significance of a red traffic light*"; "*the signification of Chinese characters*"; "*the import of his announcement was ambiguous*"
 - direct hyponym / full hyponym
 - direct hypernym / inherited hypernym / sister term
 - S: (n) message, content, subject matter, substance (what a communication that is about something is about)
 - S: (n) communication (something that is communicated by or to or between people or groups)
 - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
 - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

The Wordnet hierarchy & the synsets

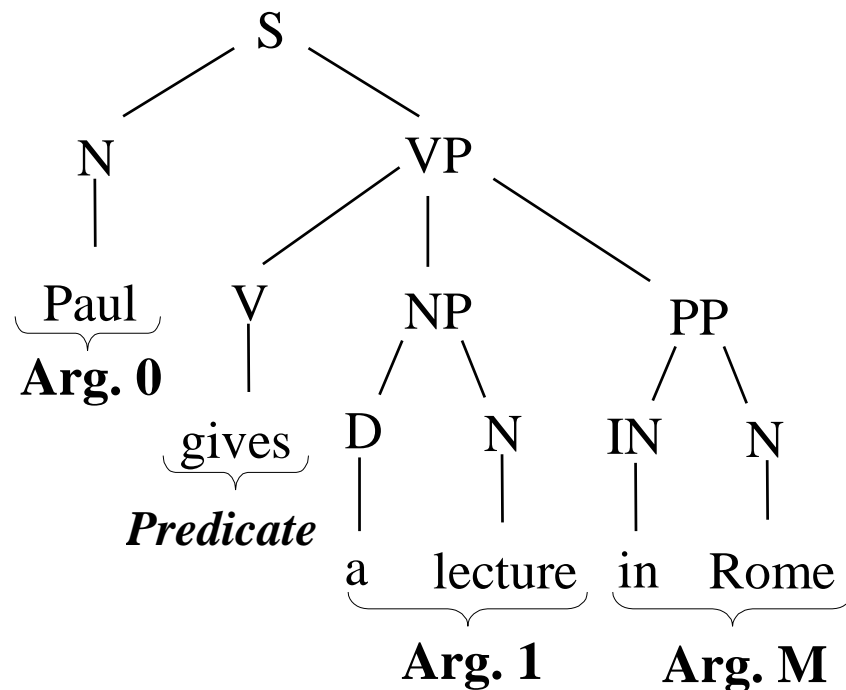


Formal, or sentential, Semantics

- Goal: Using lexical semantics and formal semantics to provide a meaning representation formalism to entire sentences
- **Semantic Parsing**: usually the process to build the formal semantic representation (of the meaning) of a sentence s using s and its (possibly multiple) grammatical representations (i.e. a parse tree or a dependency graph) as input.
- In Semantic Parsing the emphasis is the Computational aspects such as:
 - **Complexity** of the parsing process
 - **Sustainability** of the maintenance of the large lexical and ontological KBs involved
 - **Learnability** of the involved resources (e.g. lexical preferences, semantic similarity metrics, ...)
- A crucial aspect in sentential semantics is the syntax-semantics mapping required to interpret individual grammatical structures into formal logic predicates

Semantic Predicates and Arguments

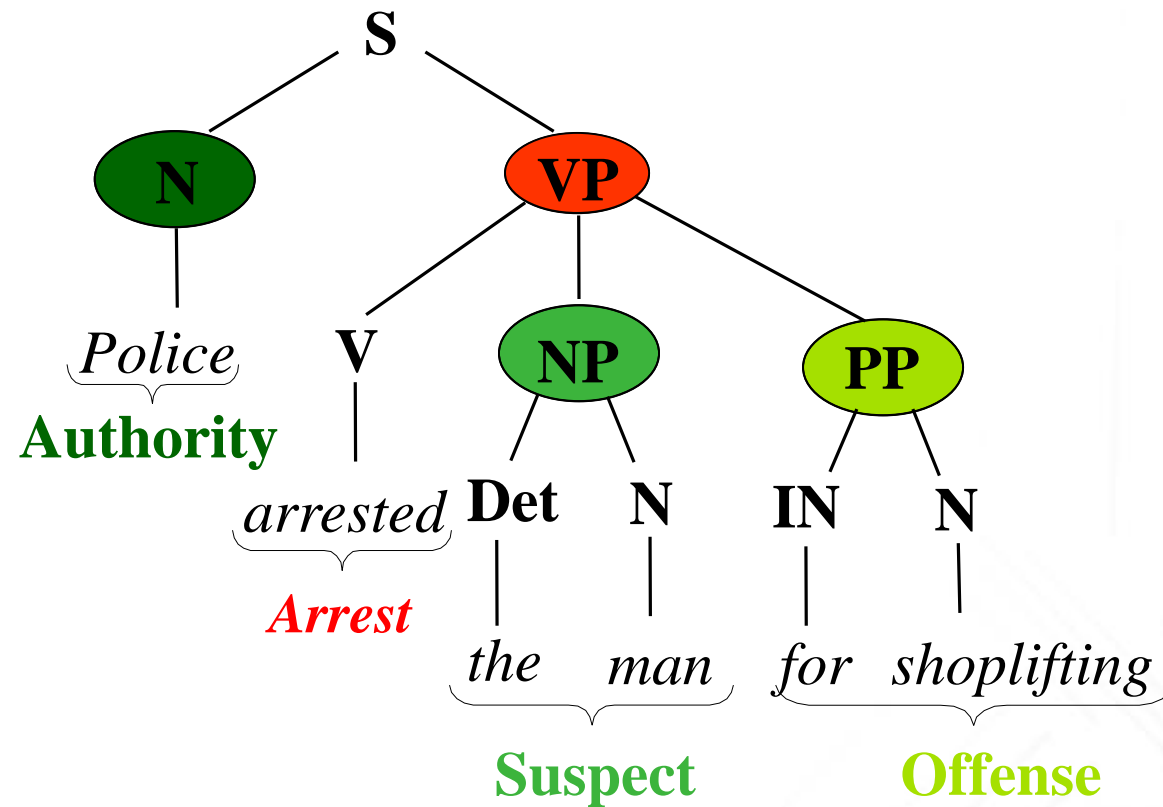
- The syntax-semantic *mapping*
- *Are there any general formalism to denote predicates?*



Different formalisms (Annotation schemes) are used as a reference model for predicates: PropBank vs. FrameNet

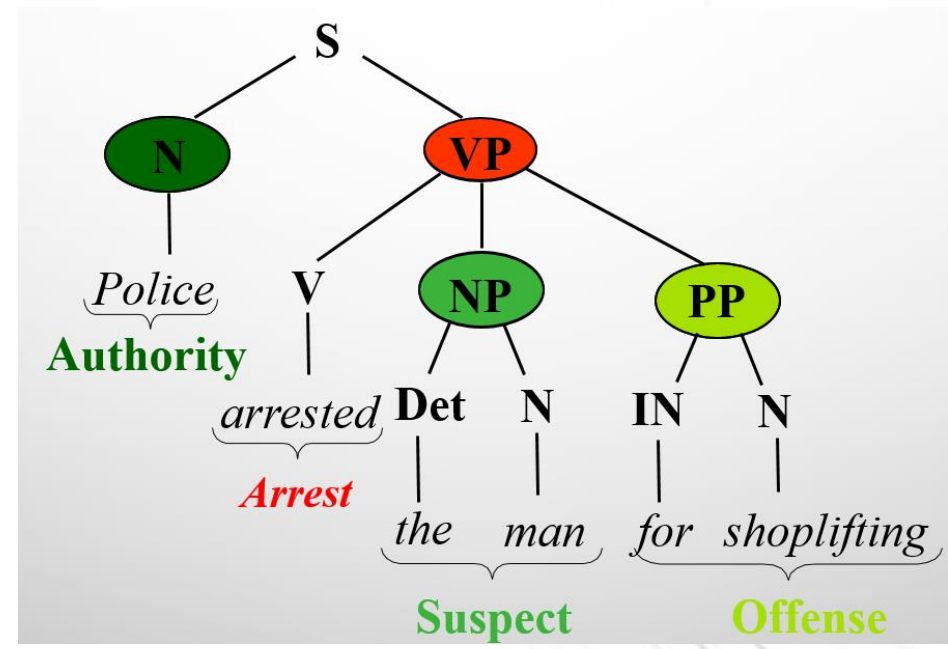
Linking syntax to semantics: the Framenet style

- *Police arrested the man for shoplifting*



FrameNet labeling: the tabular vision

Word	Predicate	Semantic Role
• <i>Police</i>	-	AUTHORITY
• <i>arrested</i>	Target	Arrest
• <i>the</i>	-	SUSPECT
• <i>man</i>	-	SUSPECT
• <i>for</i>	-	OFFENSE
• <i>shoplifting</i>	-	OFFENSE



Lexical and Sentential Semantics: Predicates & Thematic roles

- Arguments play specific roles, called *thematic roles*, depending on the predicate but invariant across different syntactic structures. They give rise to predicate argument structures
 - e.g. *Bob gives Mary the book, Bob gives the book to Mary*
are two synt. structures mapped into the invariant predicate
`give(Agent: Bob, Theme: the_book, Recipient: Mary)`
- Thematic roles of individual arguments are indexed by their predicates
 - `Agent` is the first argument of a `give/3` predicate
- Such Roles can be **general** or depend on lexical items (in this case they are called **lexicalized** roles)
 - **Agent** of a `buy/3` predicate vs. **Buyer**

THEMATIC ROLES

AGENT: Deliberately performs the action described by the verb

THEME (PATIENT): Undergoes the action of the verb or is in the state described by the verb

EXPERIENCER: Experiences the emotional or mental state or change described by the verb

INSTRUMENT: Entity used to carry out the action described by the verb

LOCATION: Place where action or state occurs

GOAL: Place toward which action is directed

SOURCE: Place from which action originates

ASSOCIATIVE: Performs action with Agent.

Frame Semantics

- Research in Empirical Semantics suggests that words represents categories of experience (situations)
- A frame is a cognitive structuring device (i.e. a kind of prototype) indexed by words and used to support understanding (Fillmore, 1975)
 - Lexical Units evoke a Frame in a sentence
- Frames are made of elements that express participants to the situation (Frame Elements)
- During communication LUs evoke the frames

Frame Semantics: KILLING

Frame: KILLING	
A KILLER or CAUSE causes the death of the VICTIM.	
Frame Elements	KILLER John <u>drowned</u> Martha.
	VICTIM John <u>drowned</u> Martha .
	MEANS The flood <u>exterminated</u> the rats by cutting off access to food .
	CAUSE The rockslide <u>killed</u> nearly half of the climbers.
INSTRUMENT It's difficult to <u>suicide</u> with only a pocketknife .	
Predicates	annihilate.v, annihilation.n, asphyxiate.v, assassin.n, assassinate.v, assassination.n, behead.v, beheading.n, blood-bath.n, butcher.v, butchery.n, carnage.n, crucifixion.n, crucify.v, deadly.a, decapitate.v, decapitation.n, destroy.v, dispatch.v, drown.v, eliminate.v, euthanasia.n, euthanize.v, ...

Frame Semantics

- Lexical descriptions are expected to define the indexed frame and the frame elements with their realization at the syntactic level:
 - *John bought a computer from Janice for 1000 \$*
- Mapping into syntactic arguments
 - the buyer is (usually) in the subject position
- Obligatory vs. optional arguments
- Selectional preferences
 - The seller and the buyer are usually “**humans**” or “**social groups**”

An example from Babel (SAG)

- Example

A law enforcement official told CNN that the FBI was investigating.

- VS

CNN was told that the FBI was investigating by a law enforcement official

- VS

CNN was told by a law enforcement official that the FBI was investigating

Babel output:

Telling: *[CNN]*_{Addressee} was **told** *[that the FBI was investigating by a law enforcement official]*_{Message} .

Law: CNN was told that the FBI was investigating by a **law** enforcement official .

Leadership: CNN was told that the FBI was investigating by a *[law enforcement]*_{Governed} **official** .

Scrutiny: CNN was told that *[the FBI]*_{Cognizer} was **investigating** *[by a law enforcement official]*_{Cognizer} .

Show CONLL format

Law: A **law** enforcement official told CNN that the FBI was investigating .

Leadership: A *[law enforcement]*_{Governed} **official** told CNN that the FBI was investigating .

Scrutiny: A law enforcement official told CNN that *[the FBI]*_{Cognizer} was **investigating** .

Telling: *[A law enforcement official]*_{Speaker} **told** *[CNN]*_{Addressee} *[that the FBI was investigating]*_{Message} .

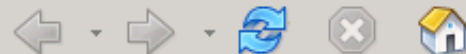
Show CONLL format

Telling: *[CNN]*_{Addressee} was **told** *[by a law enforcement official]*_{Speaker} *[that the FBI was investigating]*_{Message} .

Law: CNN was told by a **law** enforcement official that the FBI was investigating .

Scrutiny: CNN was told by a law enforcement official that *[the FBI]*_{Cognizer} was **investigating** .

Show CONLL format



Frame Report (recent data)

[| Top of Frame Index](#) | [| Top of Lexical Unit Index](#) |

Committing_crime

Definition:

A **Perpetrator** (generally intentionally) commits a **Crime**, i.e. does something not permitted by the laws of society.

They PERPETRATED a felony by substituting a lie for negotiations.

The suspect had allegedly COMMITTED the crime to gain the attention of a female celebrity.

FEs:

Core:

Crime [Cr]

An act, generally intentional, that has been formally forbidden by law.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Perpetrator [Perp] The individual that commits a **Crime**.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Non-Core:

Frequency [Freq] The frequency with which a **Crime** is committed.

The average serial killer COMMITTS a crime every five years.

Instrument [Inst] The **Instrument** used in committing the crime.

Most crimes are COMMITTED with a firearm.

Killing

D

FEs:

A

Non-Core:

F

Beneficiary [ben]

This extra-thematic FE applies to participants that derive a benefit from the occurrence of the event specified by the target predicate.

C

Circumstances []

Circumstances describe the state of the world (at a particular time and place) which is specifically independent of the event itself and any of its participants.

C

Ex

Semantic Type: Physical_entity

It's difficult to **SUICIDE** with only a pocketknife.

Excludes: Cause

Instru

Semant

Exclud

Killer [Kill]

The person or sentient entity that causes the death of the **Victim**.

Excludes: Cause

Killer

Means []

The method or action that the **Killer** or **Cause** performs resulting in the death of the **Victim**.

Exclud

Semantic Type: State_of_affairs

The flood **EXTERMINATED** the rats by cutting off access to food.

Mean

Excludes: Cause

Semant

Exclud

Victim []

The living entity that dies as a result of the killing.

Victim

Semant

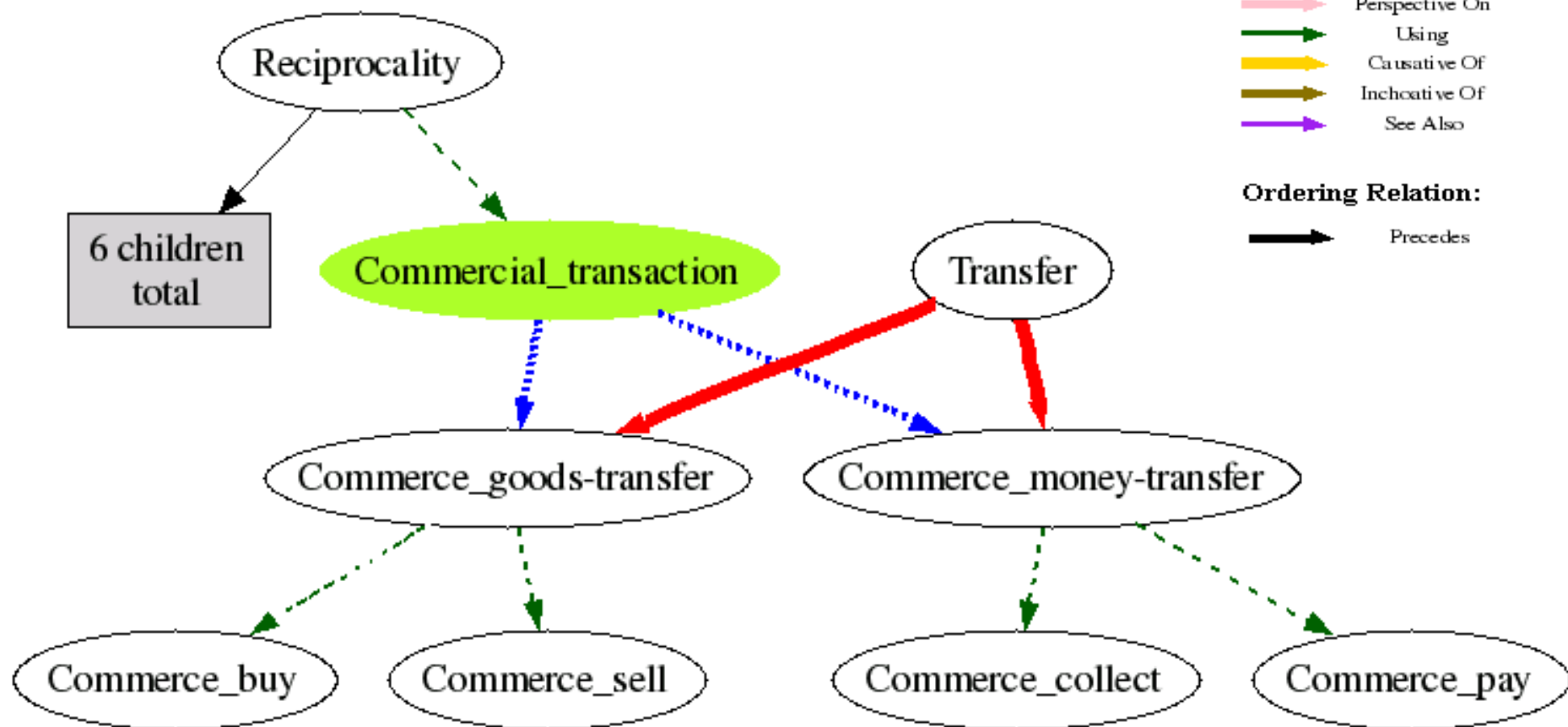
Semantic Type: Sentient

Non-Core:

Beneficiary [ben]

This extra-thematic FE applies to participants that derive a benefit from the occurrence of the event specified by the target predicate.

The FrameNet Hierarchy

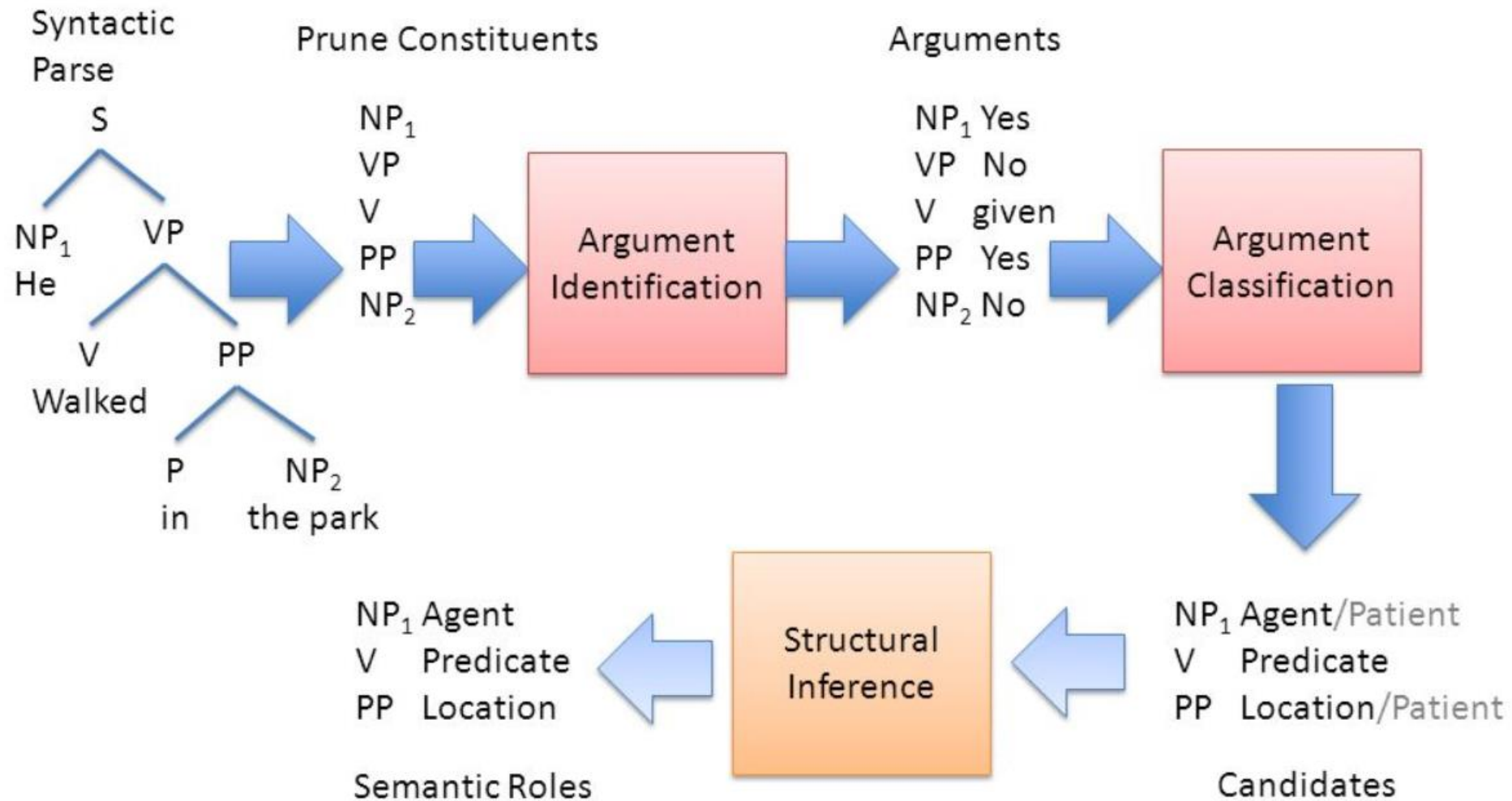


FrameNet - Data

- Methodology of constructing FrameNet
 - Define/discover/describe frames
 - Decide the participants (frame elements)
 - List lexical units that evoke the frame
 - Find example sentences in the BNC and annotate them
- Corpora
 - FrameNet I -British National Corpus only
 - FrameNet II -LDC North American Newswire corpora
- Size
 - >10,000 lexical units, >825 frames, >135,000 sentences
- <http://framenet.icsi.berkeley.edu>

Machine Learning over Framenet/PropBank

SRL Pipeline



Frame Semantics

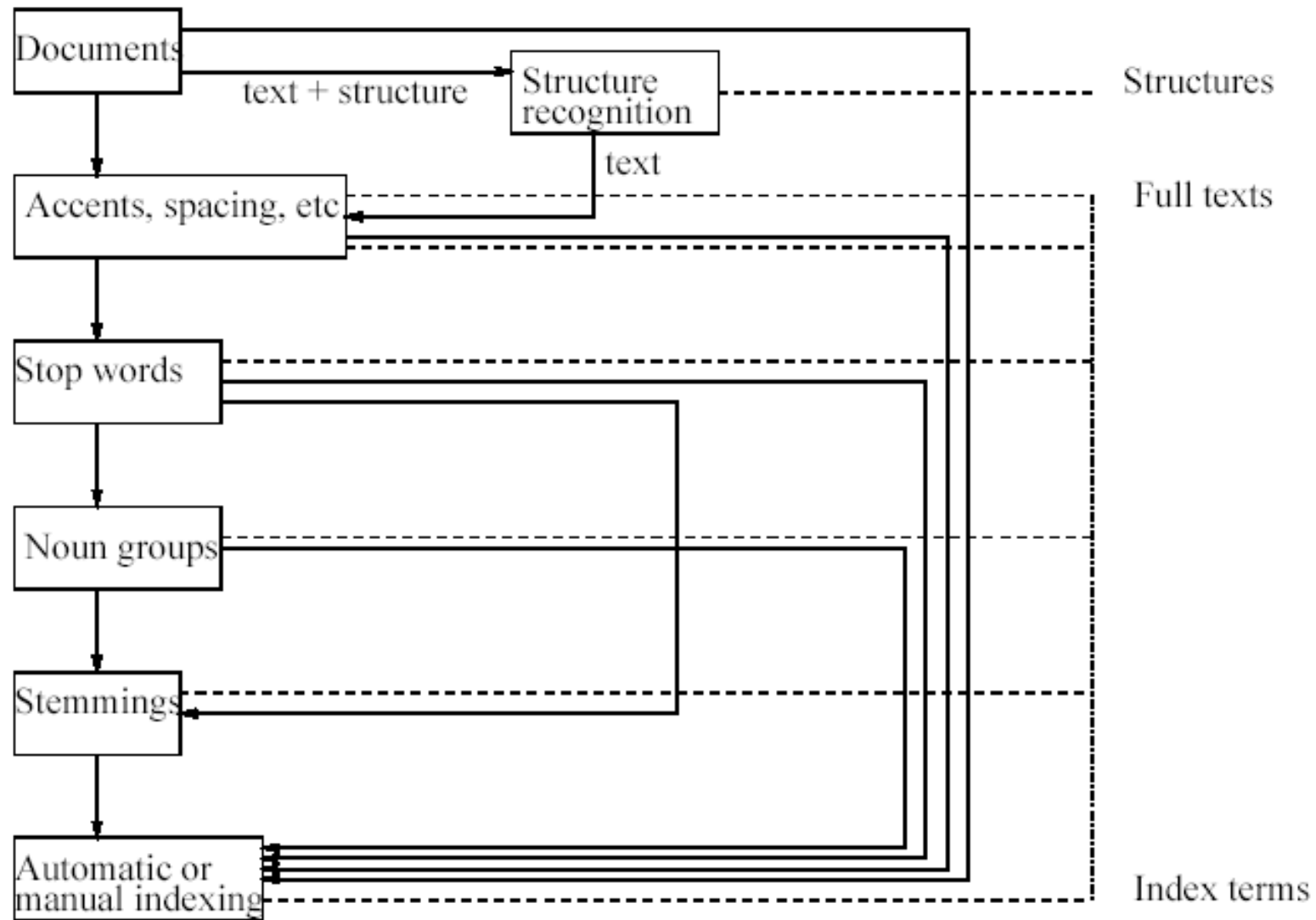
- [Charles J Fillmore](#). 1968. [The case for case](#). In [E Bach](#) and [Harms, R](#), *Universals in Linguistic Theory*, Universals in Linguistic Theory. Holt, Rinehart & Winston, New York, edition. [Google Scholar](#), [BibTex](#), [Tagged](#), [XML](#), [RIS](#)
- [Charles J Fillmore](#). 1976. [Frame semantics and the nature of language](#). *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20-32. [Google Scholar](#), [BibTex](#), [Tagged](#), [XML](#), [RIS](#)
- [Charles J Fillmore](#). 2002. [Linking Sense to Syntax in FrameNet](#). In *Proceedings of 19th International Conference on Computational Linguistics*, Taipei. COLING. [Google Scholar](#)
- [Charles J Fillmore](#). 1982. [Frame semantics](#). In *Linguistics in the Morning Calm*, Linguistics in the Morning Calm. Hanshin Publishing Co., Seoul, South Korea, edition. [Google Scholar](#)
- [Collin F Baker](#), [Fillmore, Charles J](#), and [Lowe, John B](#). 1998. [The Berkeley FrameNet project](#). In *COLING-ACL '98: Proceedings of the Conference*, Montreal, Canada. [Google Scholar](#)

Overview

- Documents in Information Retrieval
 - Information, Representation, (re)current challenges, success(and unsuccess)ful stories
- Information and Content
 - Natural Language Processing: introduction to the linguistic background
 - Natural Language and Content
 - NL Syntax
 - NL Semantics
 - Document Representation and IR models
- Summary



Typical pre-processing activities in Document-based IR



Pre-Processing steps

- Removal of special characters or meaningless separators (e.g. HTML tags, punctuation, ...)
- Segmentation of the incoming text into a sequence of tokens (usually driven by spaces)
- Token Stemming to obtain simpler word forms (aprox. “word roots”)
 - *computational* → *comput*
- Pruning of irrelevant words (also called stopwords), e.g. *the, a, he, ...*)

Pre-Processing (2)

- Recognition of non compositional expressions
 - Common expressions (e.g. *being-in-a-hurry, for sake of, it rains cats and dogs*, ...)
 - Jargon (e.g. operate, put the computer to sleep, cross-validate, ...)
 - Technical terminology (e.g. *decision tree, blind model*, ...)
- Compilation of the inverted index :
from keywords to the documents they appear in

Applications: Target Semantic Phenomena



- **Entities.** Entities cited in texts (people, locations, organizations, date, numerical or monetary expressions)



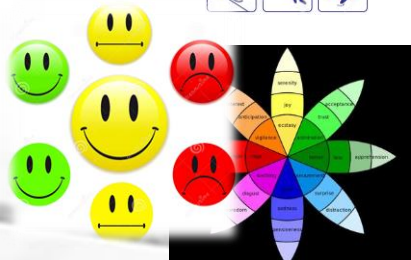
- **Relations.** Relationships / Associations among entities



- **Facts.** Facts and Events



- **Topics.** Discussion topics / Context / Domain



- **Emotional and Psychological traits.** Social Science, Profiling

NLP Applications: a roadmap



NLP over textual data

- Use of linguistic models for the **recognition of grammatical and semantic phenomena**
- **Resolution** of the main **sense ambiguities**
- **Coverage** of the involved **document sources**



Conceptualization

- **Recognition of implicit phenomena of interest**
- **Analysis of documental sources and individual fact checking**
- Discovery of **novel (global) facts** of interest



Exploration & Prediction

- **Logical checking of individual facts**
- **Aggregation of correlated facts**
- **Empirical validation of interpretation hypothesis**
Planning of more in depth analysis



Operational Knowledge and
Fact verification & Truth checking



Summary

- IR models necessary in Web mining depend on the ways unstructured data can be made available for representing texts in ML tasks such as filtering, classification, ad hoc retrieval and other ranking (e.g. recommending) tasks
- A semantic model for the content of unstructured data is strongly dependent on the linguistic nature of these latter
 - Facts, Entities, Relations, Thematic areas, Subjective information are always rooted in a form of rather free linguistic description
- Studies in Linguistics have provided the basic notions for dealing with the meaning of Natural Language expressions
 - Levels of the linguistic analysis
 - Basic paradigms: lexical description, grammars, logic as a meaning representation language

Summary (2)

- Machine Learning approaches to IR must maximize accuracy and cognitive plausibility of the decisions
- This unavoidably ask for specific models of linguistic structures such as
 - Word sets
 - Word sequences
 - Structured Texts and dialogues
 - Grammatical Trees
 - Semantic Trees and Graphs
- Algorithms (such as Nave Bayes or Rocchio's style classifiers) must be extended towards models that account for such structures in a cognitively plausible way. They **MUST** maximize both aspects of a decision:
 - **Accuracy** (What to do against some linguistic input)
 - **Epistemological transparency** (Why to do that)

Terminology

- Morphology, POS tag, Morphological derivation, root, lemma, morphological features
- Grammar, Rule, Linguistic Patterns, Derivation Trees, Dependency Graphs, Constituent, Dependency link/arc,
- Lexicon, Lexical grammatical categories, Lexical Semantics
- Computational Semantics, Logical Form, Lambda-expression
- Word sense, Frame semantics, Lexical Unit, Frame Element
- Named-Entity Recognition, Parsing, Semantic Role Labeling

Reference Textbook material

- «Speech and Language Processing”, D. Jurafsky and J. H .Martin, Prentice-Hall, 3d Edition. URL: <https://web.stanford.edu/~jurafsky/slp3/>
 - Syntax: Chapt. 12.1-12.3, 15.1-15.2
 - Semantics: 16.1-16.2, 19.1-19.3
 - Word senses: 20.1-20.3,
 - Framenet: 20.5

References

- AI & Robotics. «Robot Futures», Ilah Reza Nourbakhsh, MIT Press, 2013
- NLP & ML:
 - «Statistical Methods for Speech Recognition», F. Jelinek, MIT Press, 1998
 - «Speech and Language Processing», D. Jurafsky and J. H. Martin, Prentice-Hall, 2009.
 - «Foundations of Statistical Natural Language Processing, Manning & Schütze, MIT Press 2001.
- Sitografia:
 - SAG, Univ. Roma Tor Vergata: <http://sag.art.uniroma2.it/>
 - Reveal s.r.l.: <http://www.revealsrl.it/>

