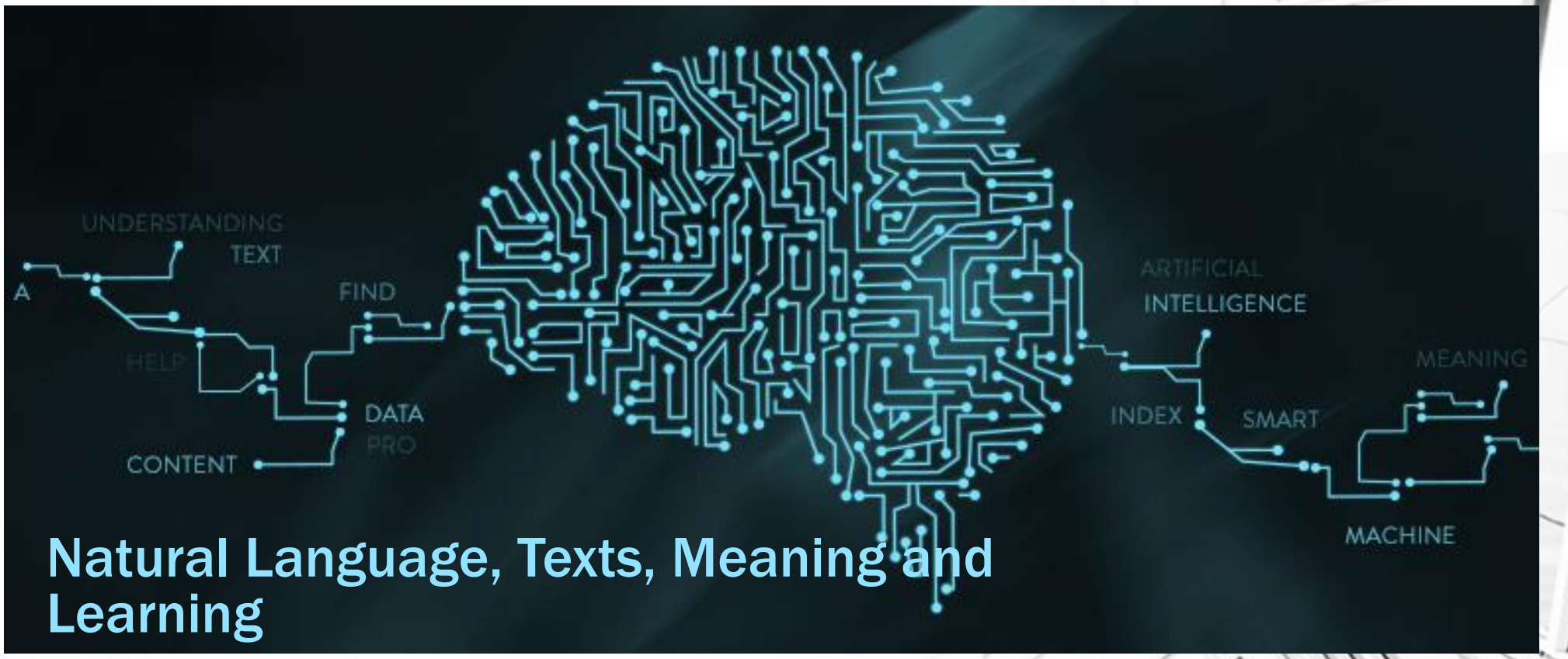


Natural Language, Texts, Meaning and Learning

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. 2022-23

Overview

- Information and Content
  - Information, Representation, (re)current challenges, success(and unsuccess)ful stories
  - Natural Language Processing: introduction to the linguistic background
    - Natural Language and Content
    - NL Syntax
    - NL Semantics
  - Language and Learning
    - Experience and Learning for NLP
    - Some Examples
- Summary

Semantics, Open Data and Natural Language

- Web contents, characterized by rich multimedia information, are mostly opaque from a semantic standpoint

Information, Web and Natural Languages

Hu meets KMT honorary chairman...

Indietro

DownloadHelper

english.peopledaily.com.cn/90785/7642916.html

人民网 English
People's Daily Online

Set as Homepage | Register | Sign In

Chinese | Big 5 | French | Russian | Spanish | Japanese | Arabic

Special Coverage | PD Online Database | Sina Microblog | Tencent Microblog | Feedback | RSS

GO

News | Opinions | China Military | Foreign Affairs | Learn Chinese | China Study | Forum

Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011. (Xinhua/Huang Jingwen)

HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central

Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011. (Xinhua/Huang Jingwen)

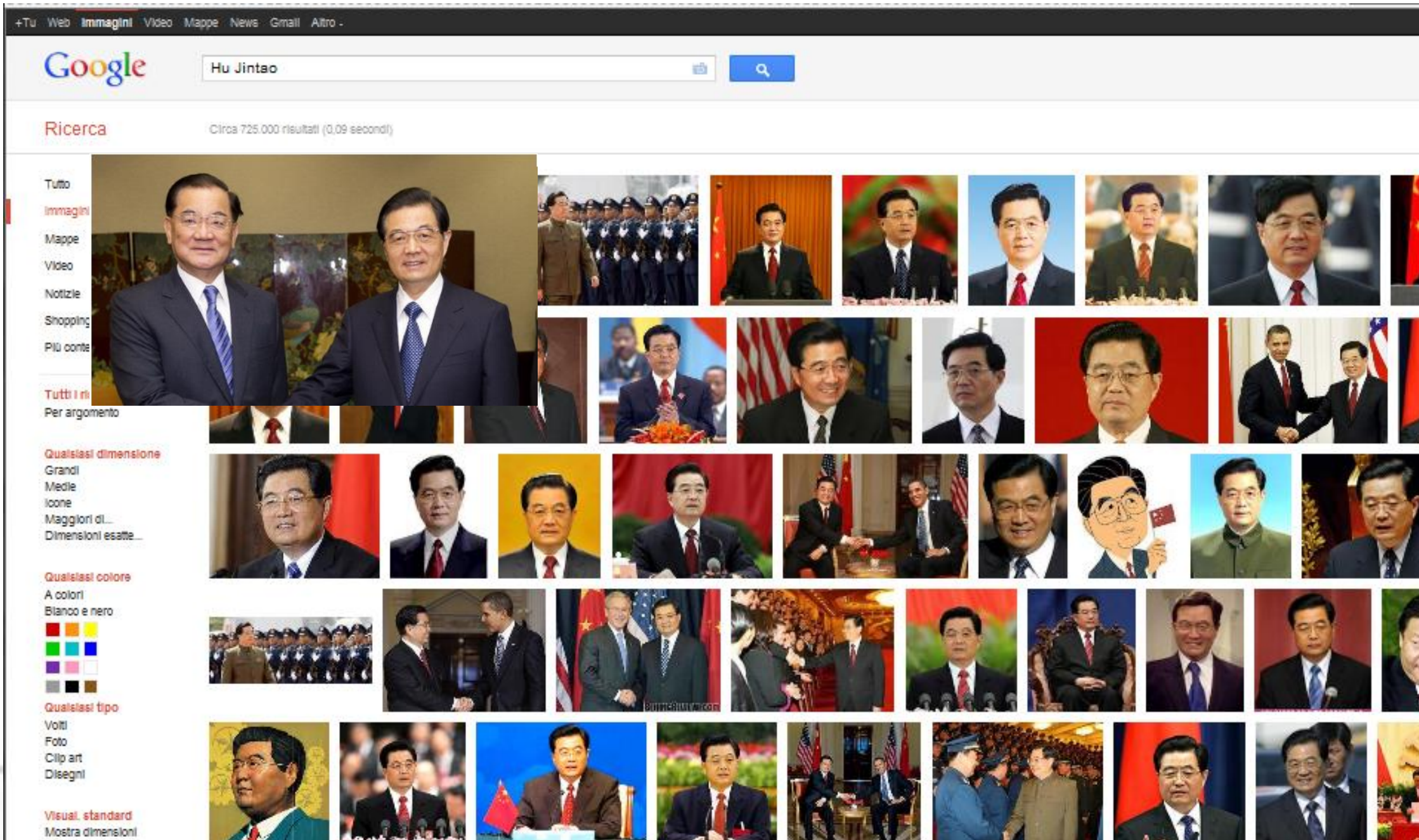
HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central

Year in SW China

Helsinki

Who is Hu Jintao?

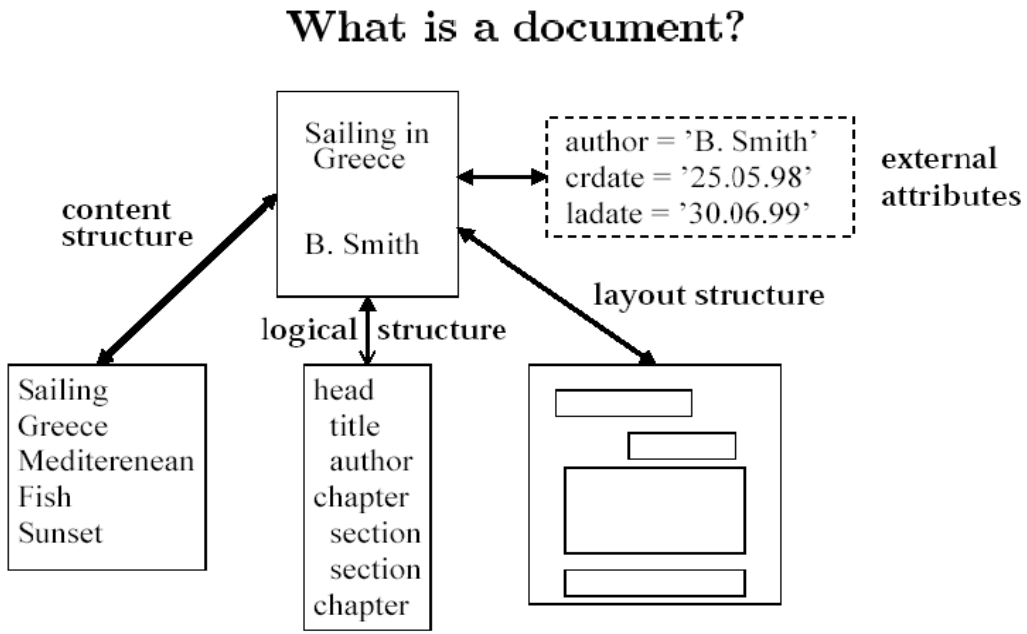
3 China in APEC: a mutually beneficial en...
4 Night life in Shanghai
5 China's 2011 foreign trade to grow 20 p...
6 Beijing house prices stumble 5.1 pct as...
7 Lama students start school in Tibet Col...
8 Police in central China crack phoney ca...
9 China-ASEAN cooperation sees notable pr...
10 Miao ethnic group celebrates Miao's New...

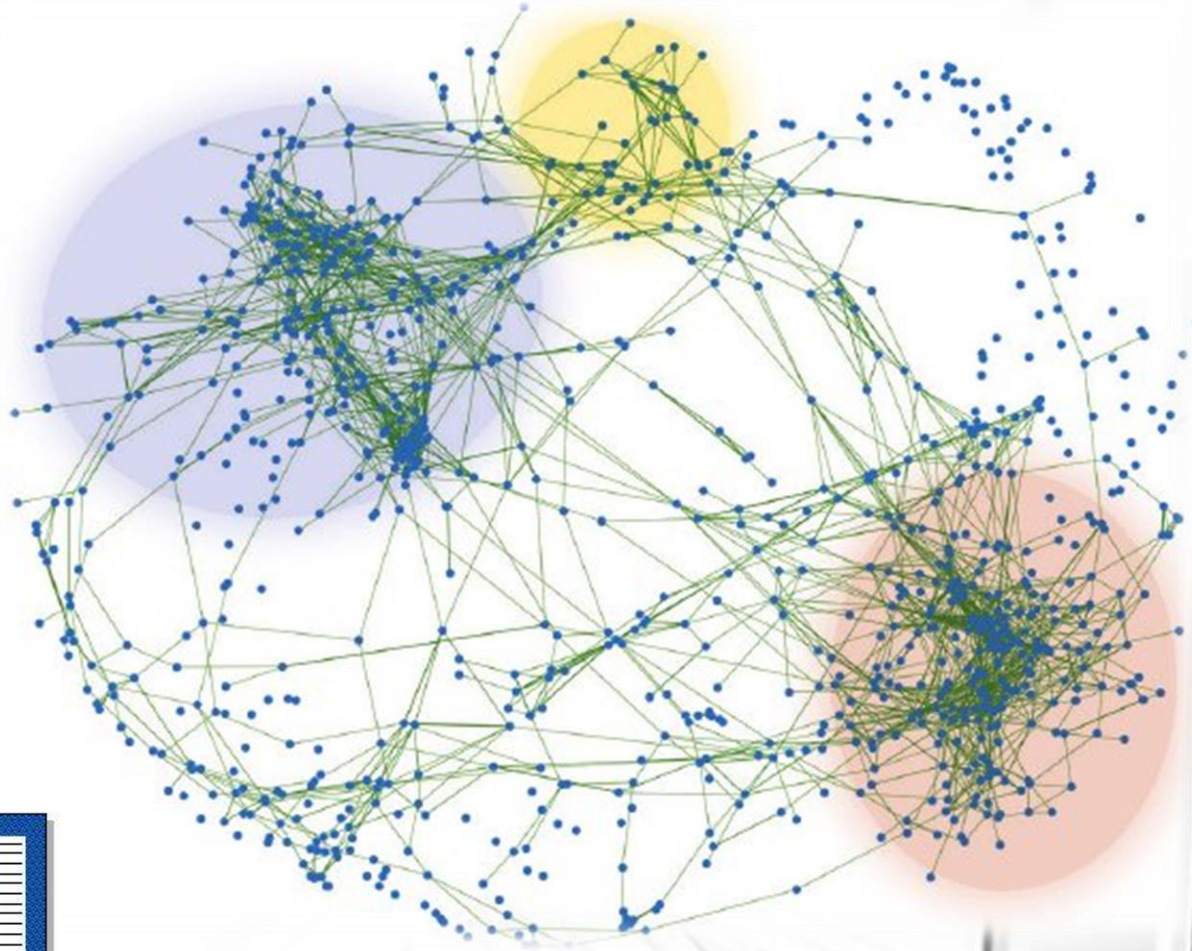

+Tu Web Immagini Video Mappe News Gmail Altro -
Google
Hu Jintao
Ricerca
Circa 725.000 risultati (0,09 secondi)
Tutto
Immagini
Mappe
Video
Notizie
Shopping
Più conte
Tutti i ri
Per argomento
Qualsiasi dimensione
Grandi
Medie
Icone
Maggiori di...
Dimensioni esatte...
Qualsiasi colore
A colori
Bianco e nero
Qualsiasi tipo
Volti
Foto
Clip art
Disegni
Visual. standard
Mostra dimensioni

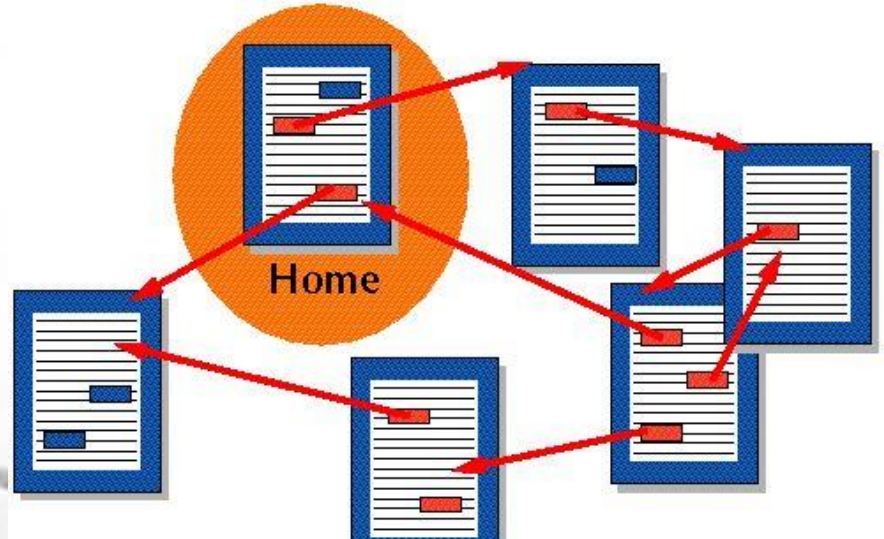
Content Semantics and Natural Language

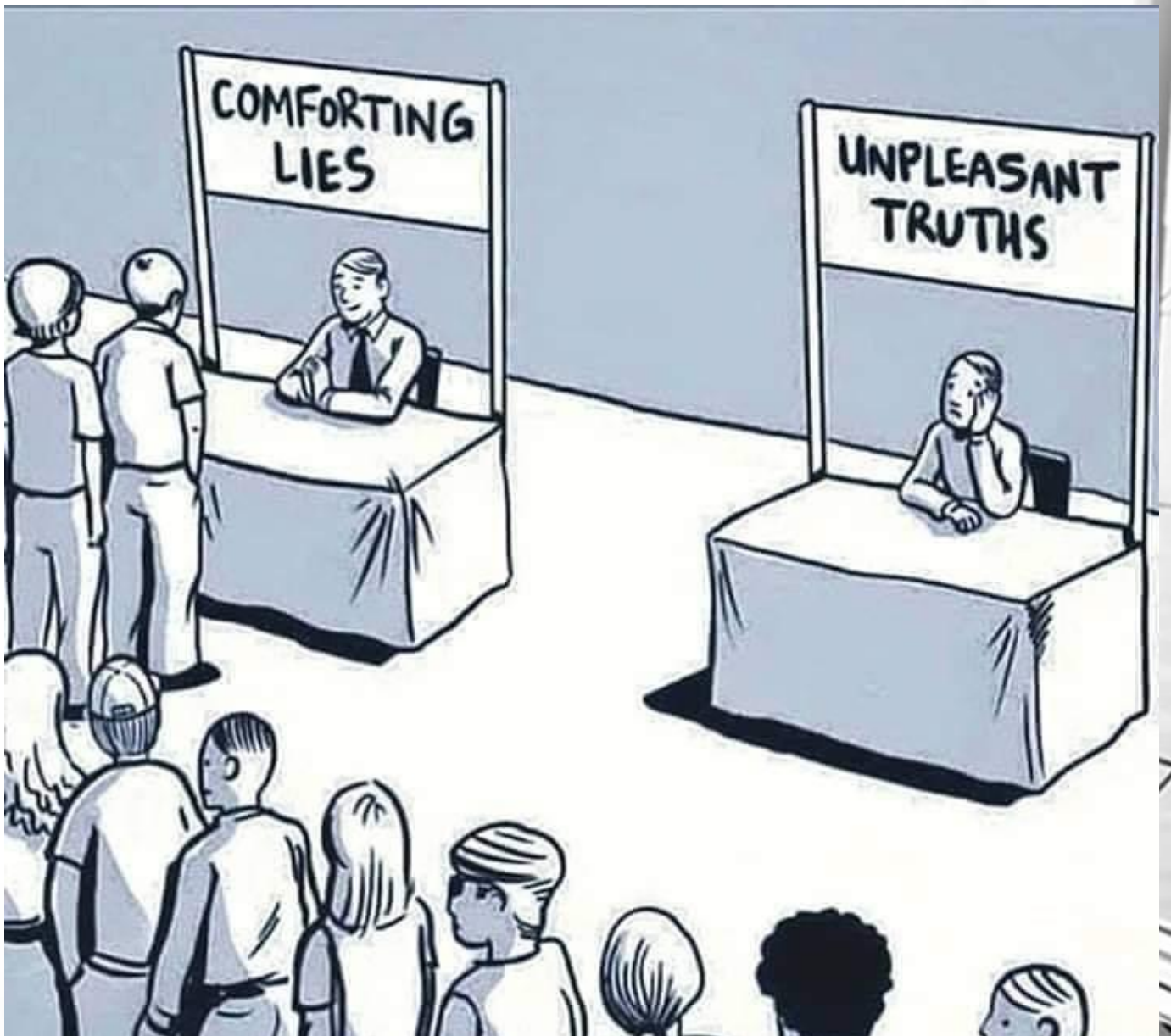
- Human languages are the main carrier of the information involved in processes such as retrieval, publication and exchange of knowledge as it is associated to the open Web contents
- Words and NL syntactic structures express concepts, activities, events, abstractions and conceptual relations we usually share through data
- “Language is parasitic to knowledge representation languages but the viceversa is not true” (Wilks, 2001)
- From Learning to Read to Knowledge Distillation as a(n integrated pool of) Semantic interpretation Task(s)

Texts, Information & Document Structures








COMFORTING LIES
UNPLEASANT TRUTHS

Two major objectives

- Discuss the **nature of content** in unstructured data within a semantic perspective over natural language
  - What constitute a useful notion of content within unstructured data collections (that are largely made of linguistic information, e.g. Web pages or infographics)
  - What is **natural language semantics** and how can we model it formally?
  - What is the **meaning of a linguistic expression**?
  - Which **structures** are emerging **as implicit components of linguistic data**?

- What is **the role of Machine Learning** that we can use against this structures in order to preserve meaning and linguistic knoweldge?
- Which **advantages** we can derive for Web Mining applications?

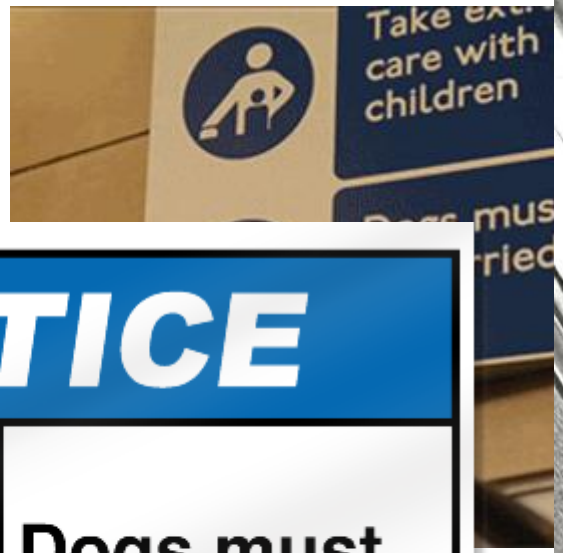
Overview

- Information and Content
  - Information, Representation, (re)current challenges, success(and unsuccess)ful stories
  - Natural Language Processing: introduction to the linguistic background
    - Natural Language and Content
    - NL Syntax
    - NL Semantics
  - Language and Learning
    - Experience and Learning for NLP
    - Some Examples
- Summary

Content in unstructured data

- Natural Language
  - Structure
  - Semantics
  - Types of semantics
  - Relationship with Machine Learning
- Examples:
  - NLU: natural language as a logic language
  - Providing more structure: Frame semantics
    - Logic, Frames and Scripts
    - The relationships between syntax and semantics
  - Semantic role labeling

Natural Language & Ambiguity



Ambiguità

- "*Dogs must be carried on this escalator*"

can be interpreted in a number of ways:

- *All dogs should have a chance to go on this wonderful escalator ride*
- *This escalator is for dog-holders only*
- *You can't carry your pet on the other escalators*
- *When riding with a pet, carry it*

The NLP chain

Levels of linguistic analyses

Pragmatics: what does it do?

Semantics: what does it mean?

Syntax: what is grammatical?

natural language utterance

Analogy with artificial languages

Syntax: no compiler errors

Semantics: no implementation bugs

Pragmatics: implemented the right algorithm

Different syntax, same semantics (5):

$$2 + 3 \quad \Leftrightarrow \quad 3 + 2$$

Same syntax, different semantics (1 and 1.5):

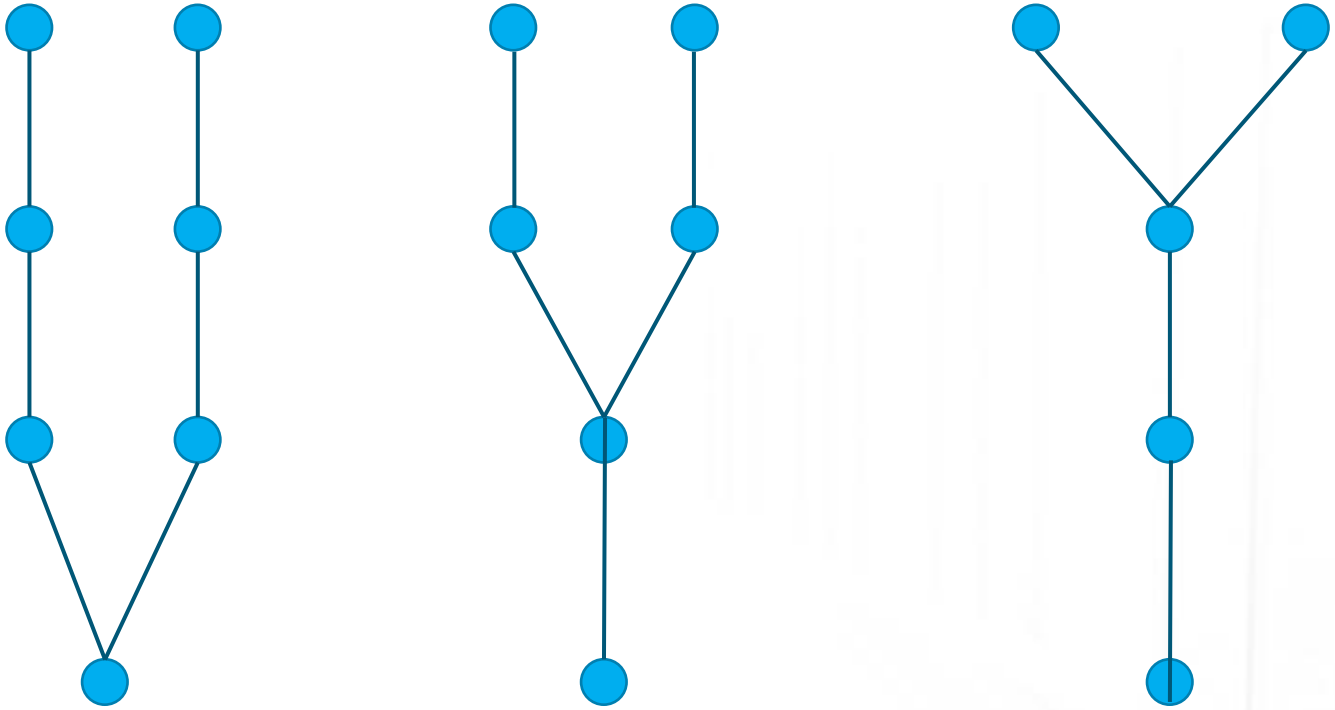
$$3 \;/\; 2 \text{ (Python 2.7)} \quad \not\Leftrightarrow \quad 3 \;/\; 2 \text{ (Python 3)}$$

Good semantics, bad pragmatics:

correct implementation of deep neural network
for estimating coin flip prob.

Ambiguity and Linguistic Levels

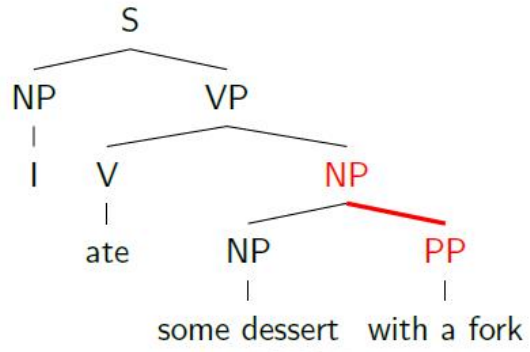
- Semantics
- Syntax
- Morphology
- Phonology

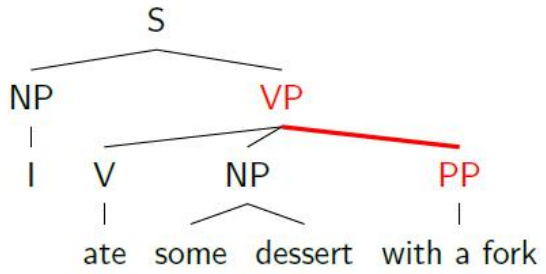


	can/can	eat cake with fork	earth observation satellite Eco's book
	del (pane) /del (libro)	compro la borsa in pelle	il timore dei manager

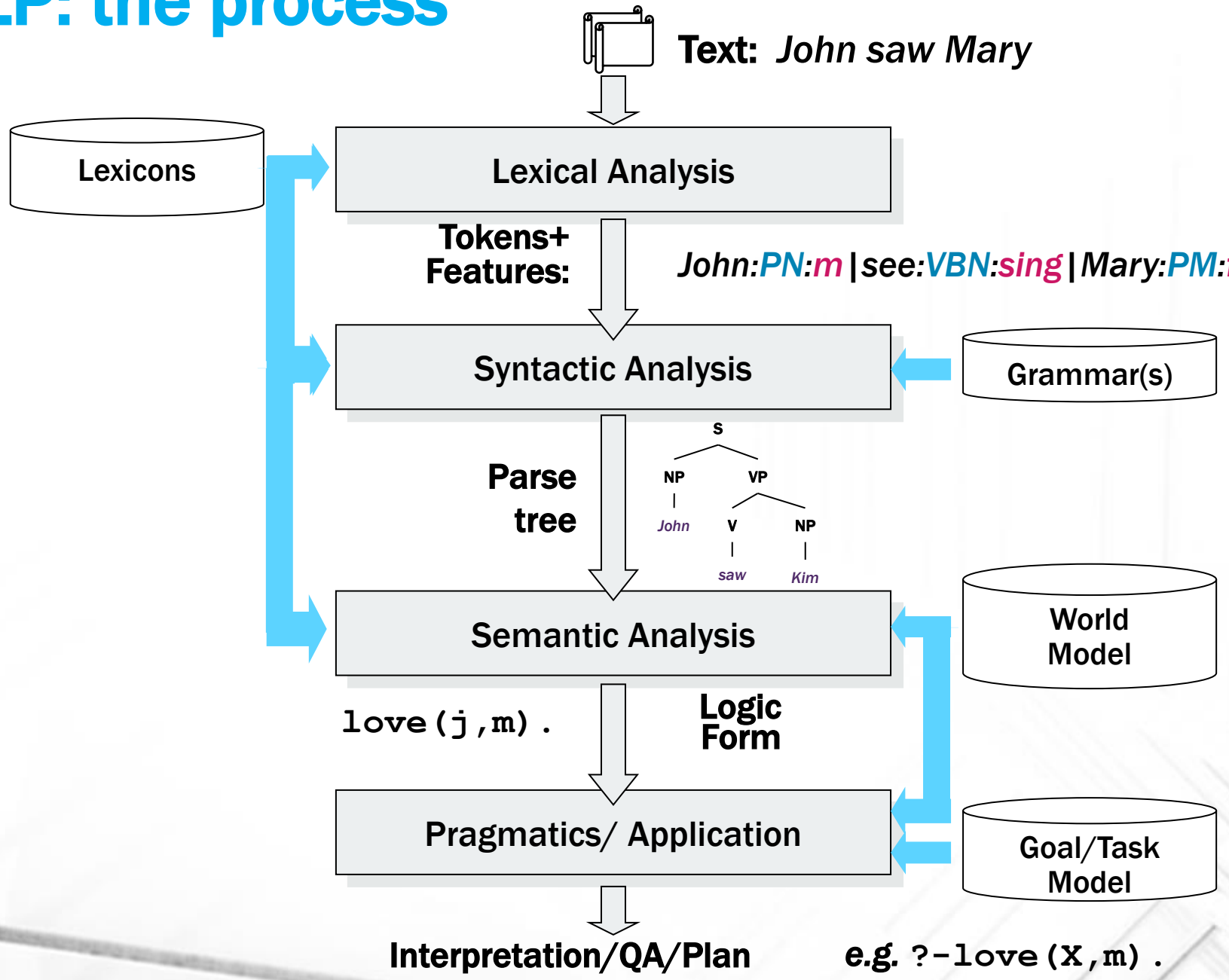
Grammars & Ambiguity

I ate some dessert with a fork.





NLP: the process



Syntax

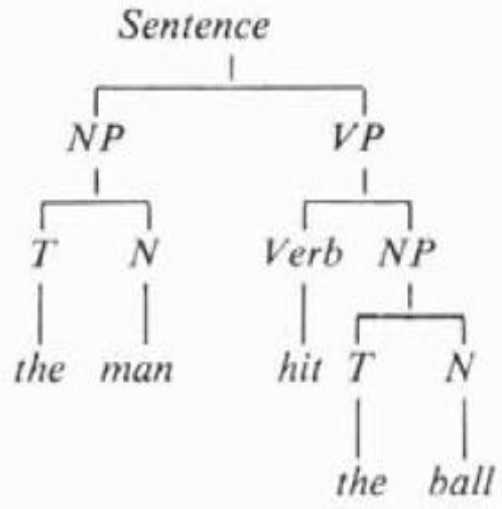
- In linguistics, syntax is *the study of the rules that govern the structure of sentences, and which determine their relative grammaticality*.
- Such rules govern a number of language phenomena as systems for phonology, morphology, syntax as well as discourse

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)

(13) (i) $Sentence \rightarrow NP + VP$
(ii) $NP \rightarrow T + N$
(iii) $VP \rightarrow Verb + NP$
(iv) $T \rightarrow the$
(v) $N \rightarrow man, ball,$ etc.
(vi) $Verb \rightarrow hit, took,$ etc.



Each such grammar is defined by a finite set Σ of initial strings and a finite set F of 'instruction formulas' of the form $X \rightarrow Y$ interpreted: "rewrite X as Y." Though X need not be a single symbol, only a single symbol of X can be rewritten in forming Y. In the grammar (13), the only member of the set Σ of initial strings was the single symbol *Sentence*, and F consisted of the rules (i) – (vi); but we might want to extend Σ to include, for example, *Declarative Sentence*, *Interrogative Sentence*, as additional symbols.

Some grammars of the form [Σ, F] may have no terminal strings, but we are interested only in grammars that do have terminal strings, i.e., that describe some language. A set of strings is called a *terminal language* if it is the set of terminal strings for some grammar [Σ, F]. Thus each such grammar defines some terminal language (perhaps the 'empty' language containing no sentences), and each terminal language is produced by some grammar of the form [Σ, F]. Given a terminal language and its grammar, we can reconstruct the phrase structure of each sentence of the language (each terminal string of the grammar) by considering the associated diagrams of the form (15), as we saw above. We can also define the grammatical relations in these languages in a formal way in terms of the associated diagrams.

Syntax: Phrase Structure Grammars (Chomsky, 75)

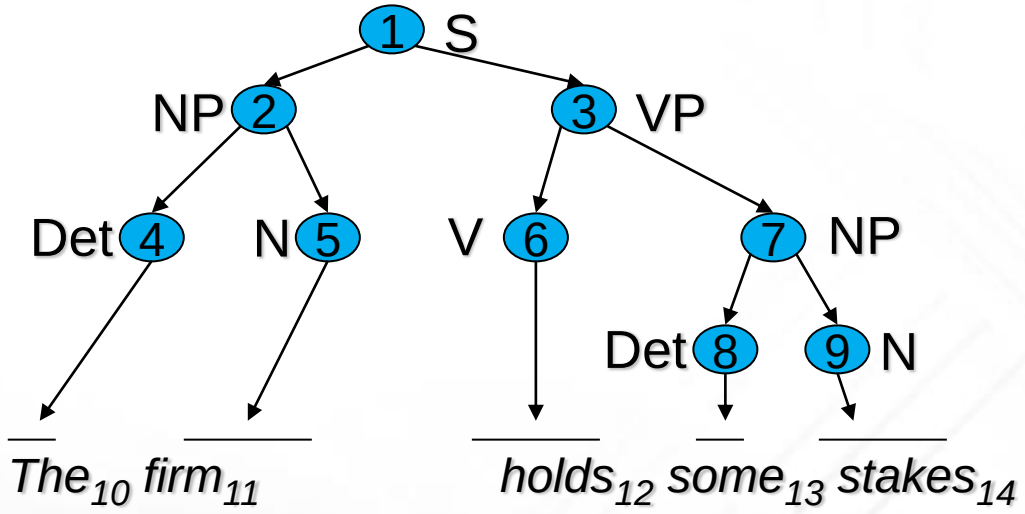
"*The firm holds some stakes*"

Symbol Vocabulary: Vn={S,NP,VP,Det,N}, VT={*The*, ...} **Axiom:** S

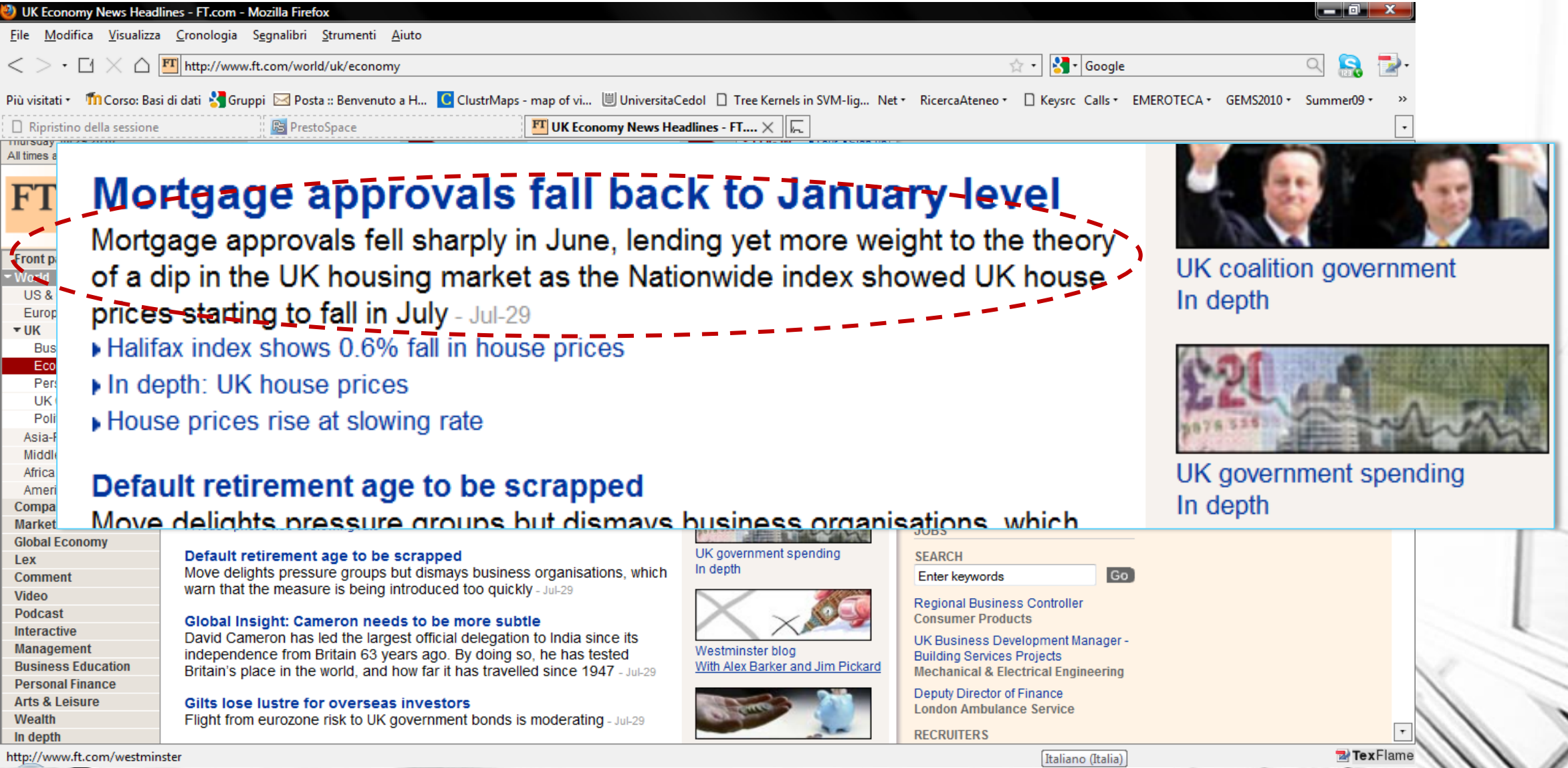
Productions: {S→NP VP, VP→V NP, NP→Det N}

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

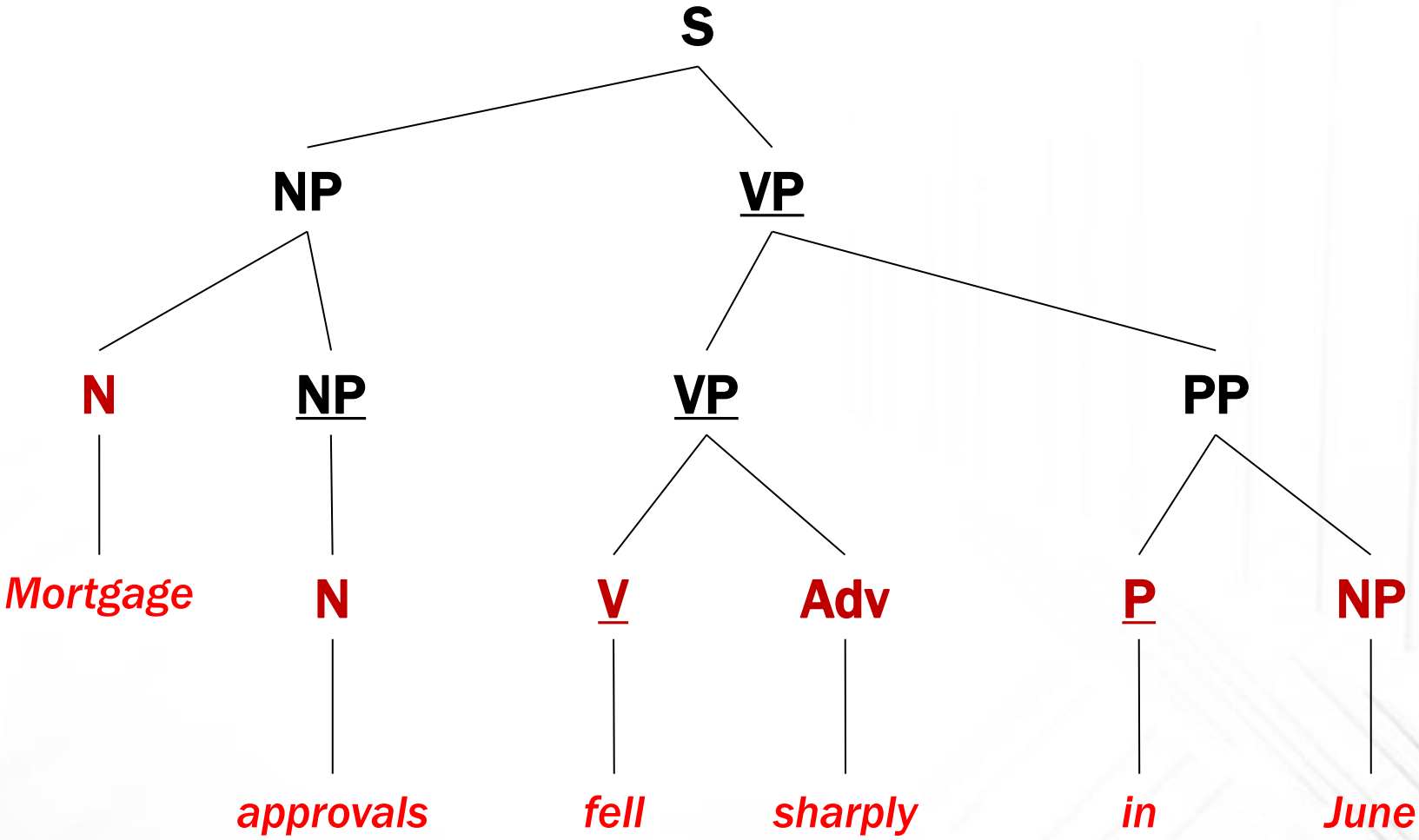
- S > NP VP > Det N VP > *The* N VP > *The firm* VP > *The firm* V NP > *The firm holds* NP > *The firm holds* Det N > *The firm holds some* N > *The firm holds some stakes*



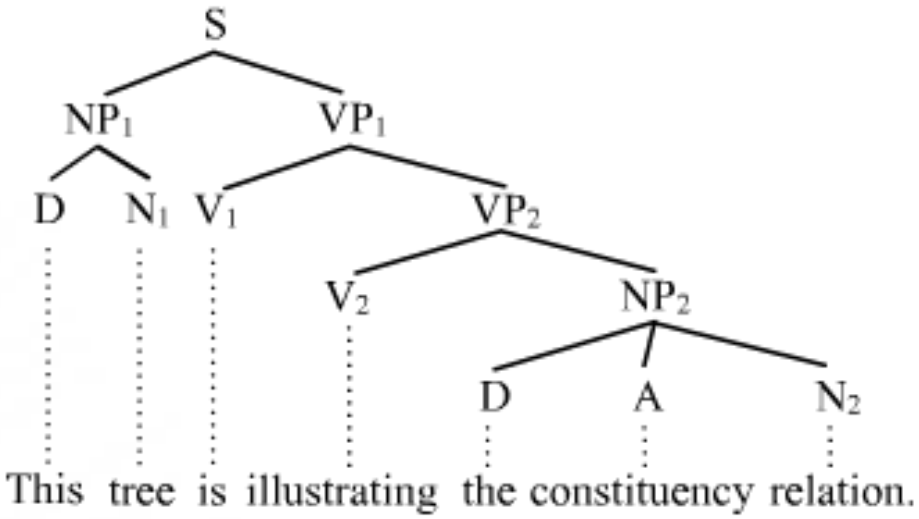
Grammatical Analysis



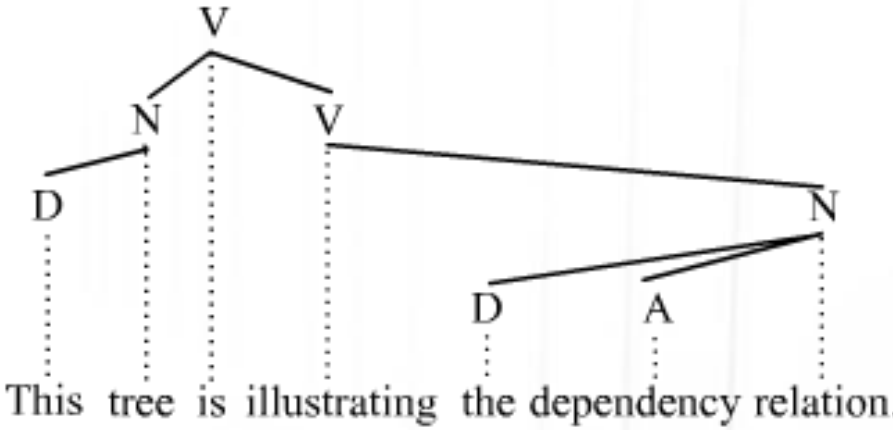
Constituent-based Parsing
(with marked Heads)



Constituency-relation vs. Dependency

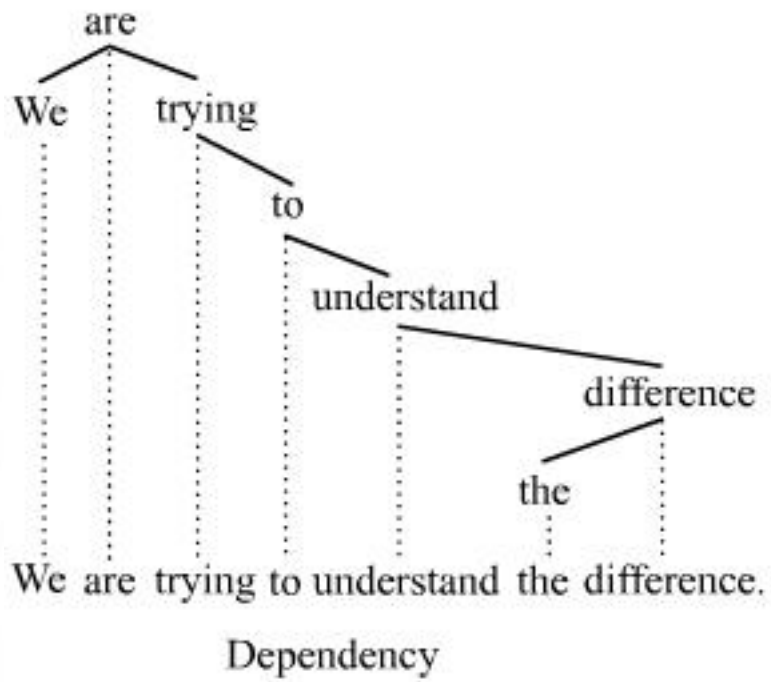


Constituency relation (PSG)

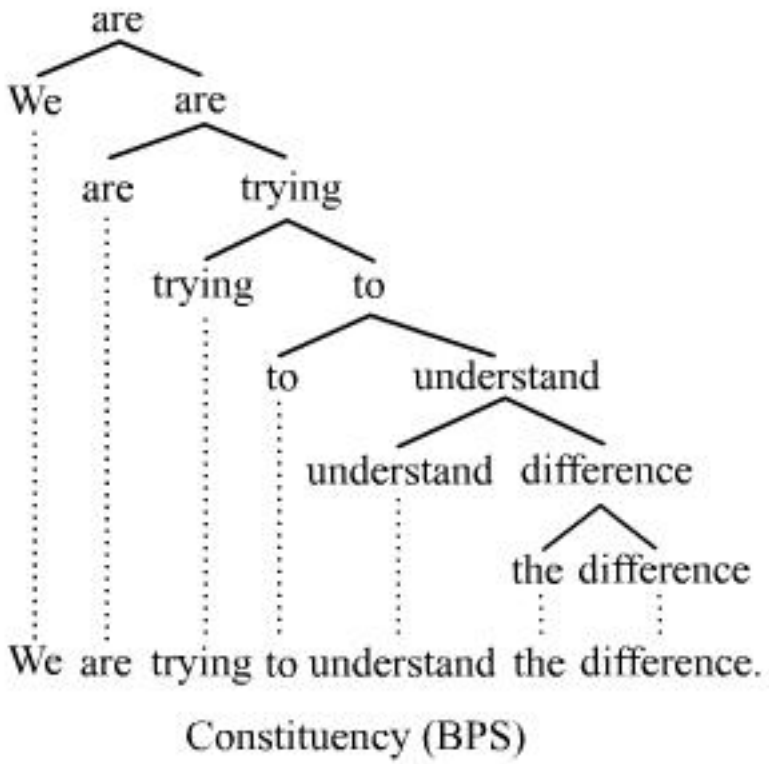


Dependency relation

Constituency vs. Dependency

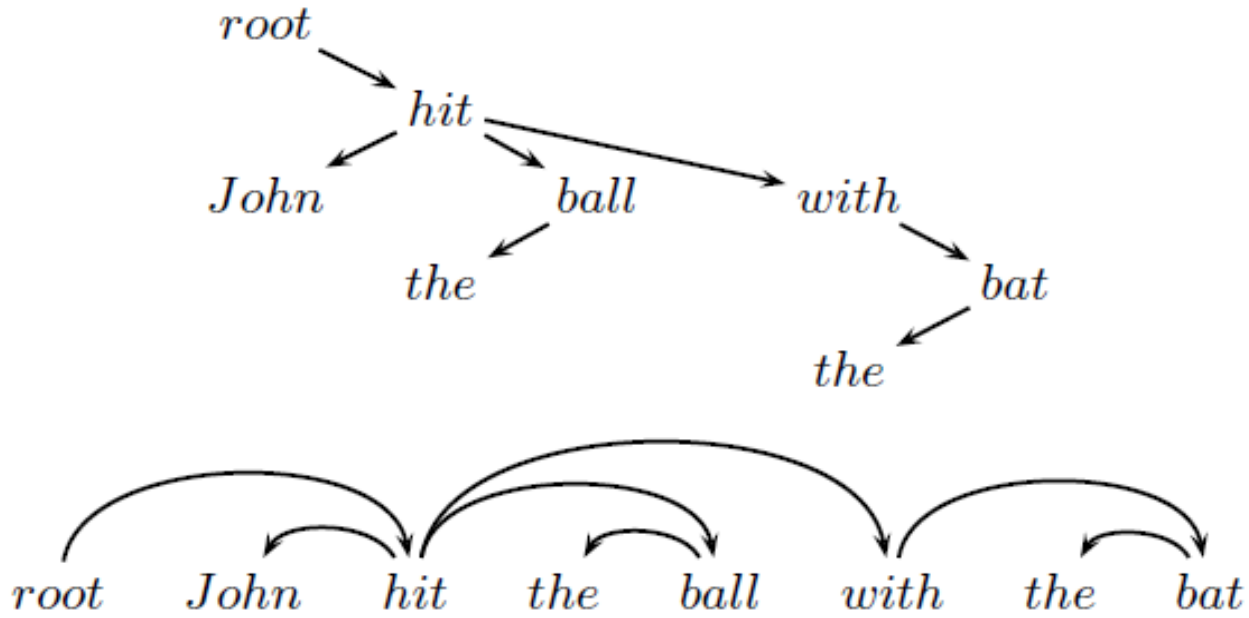


Dependency



Constituency (BPS)

From DTs to Dependency Graphs



Dependency Structures

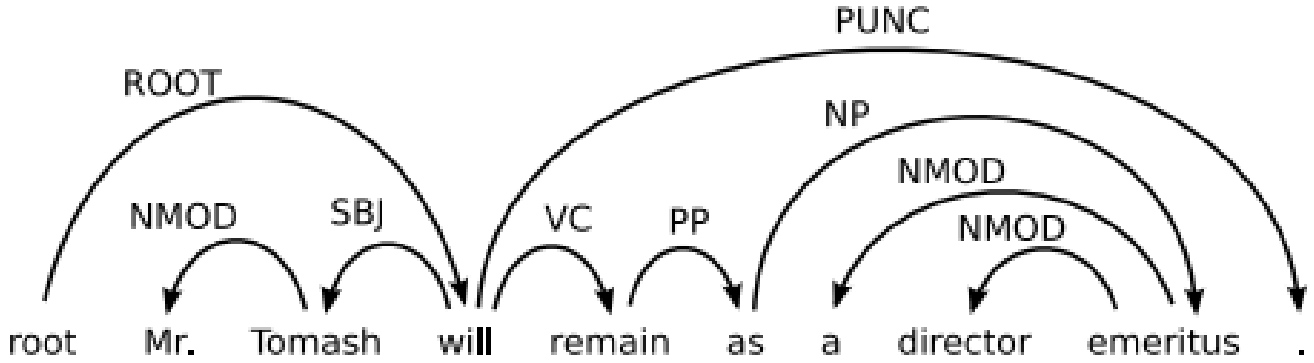


Figure 1: A projective dependency graph.

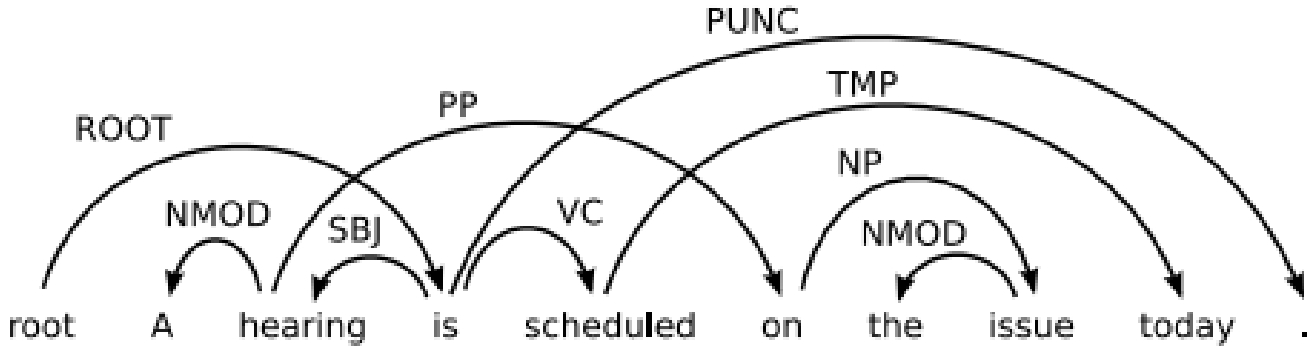
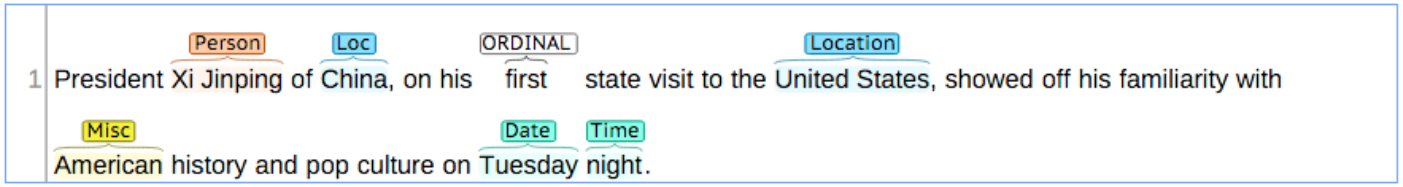


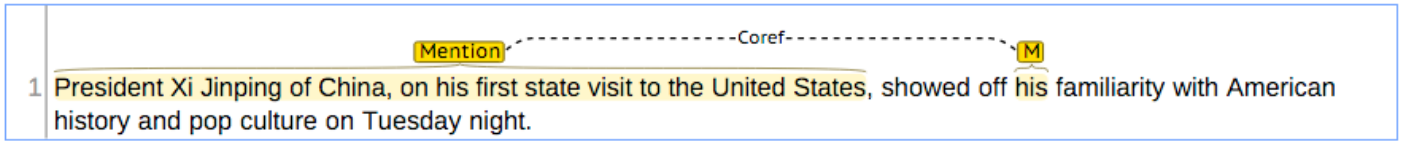
Figure 2: Non-projective dependency graph.

Dependency Parsing

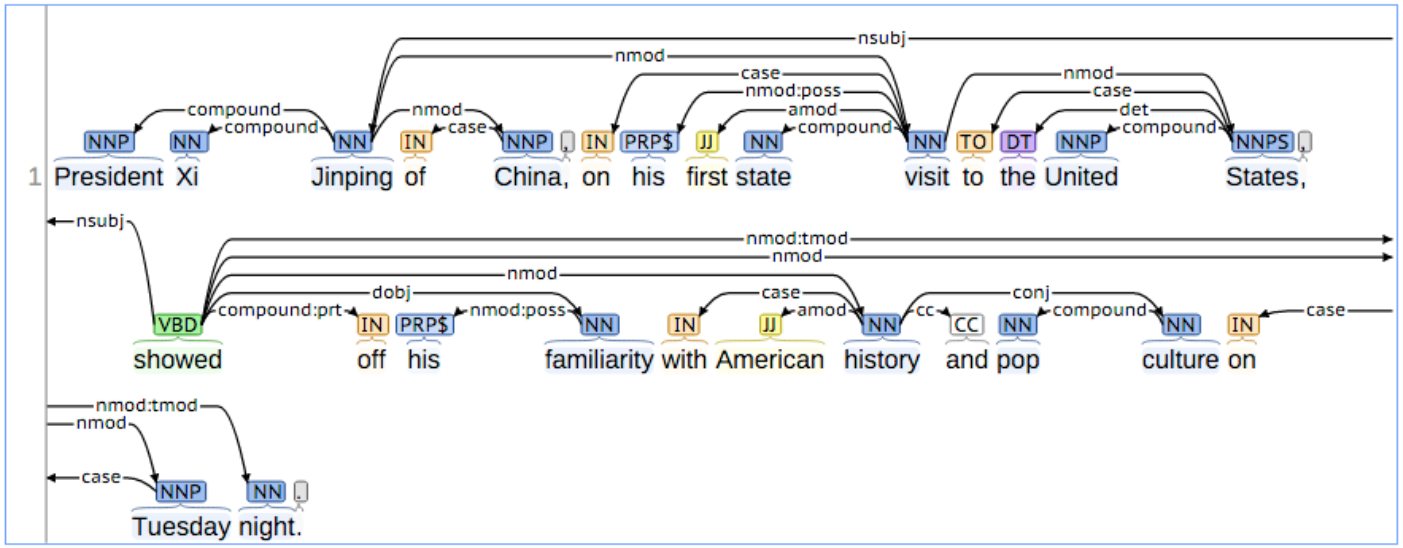
Named Entity Recognition:

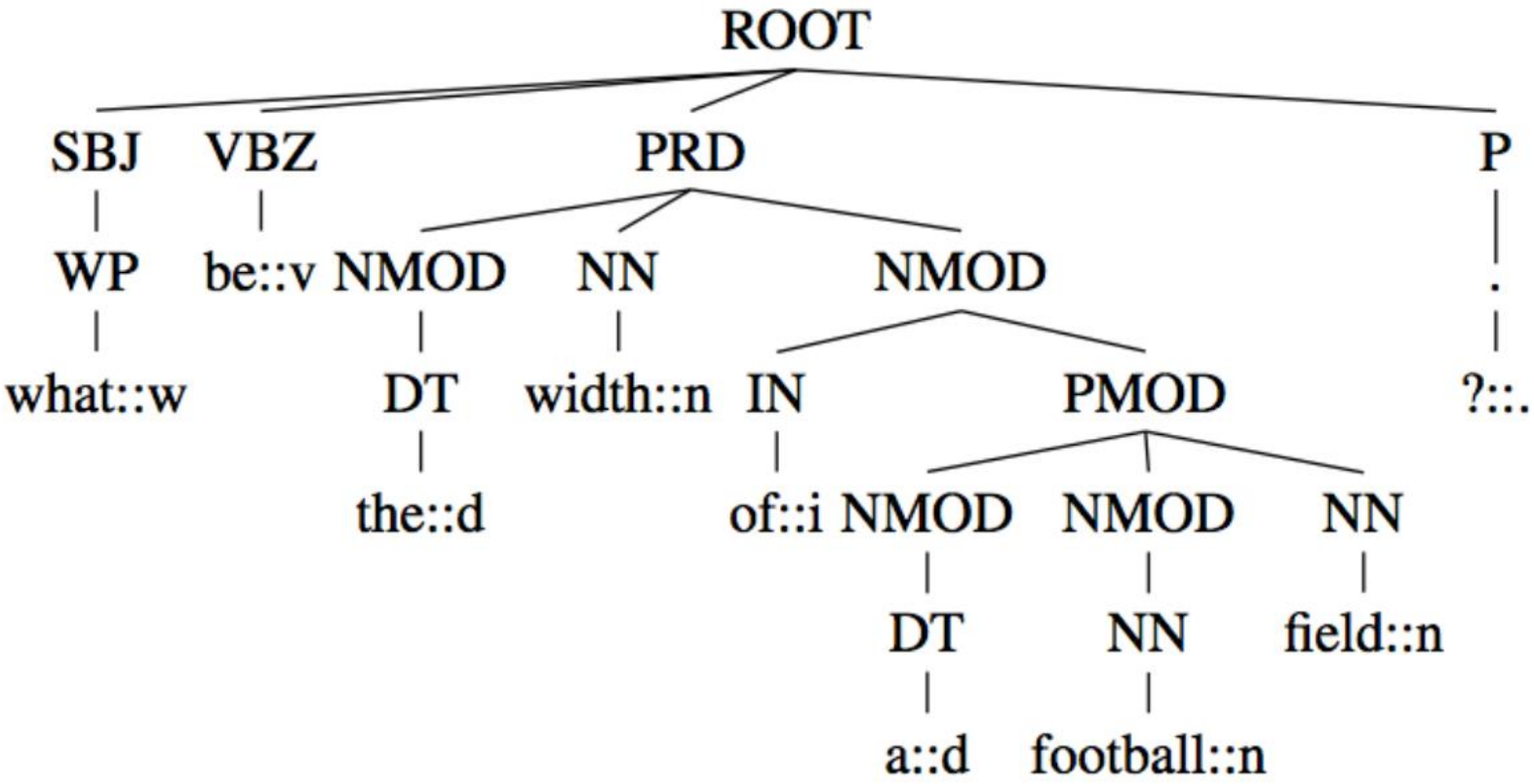


Coreference:



Basic Dependencies:

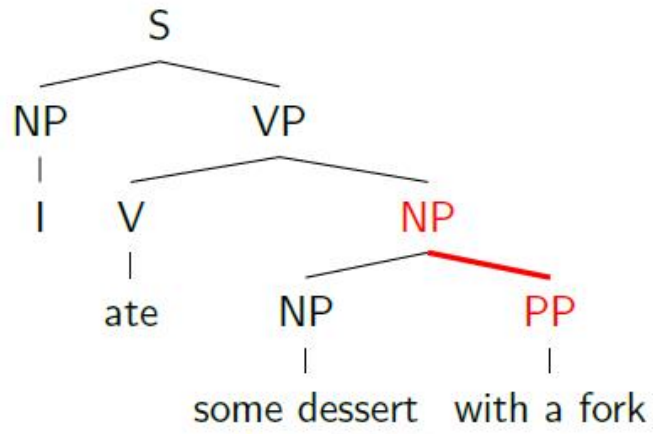


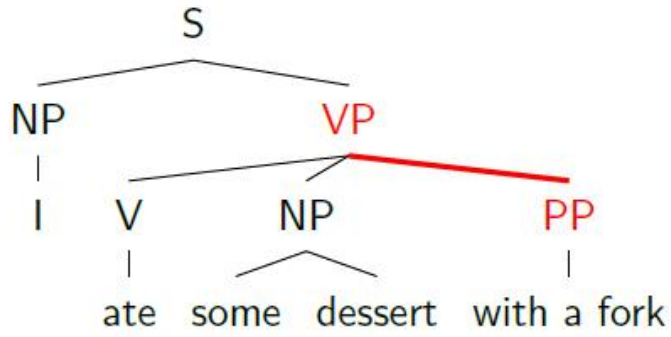


Grammatical Relation Centered Tree (GRCT)

Grammars & Ambiguity

I ate some dessert with a fork.





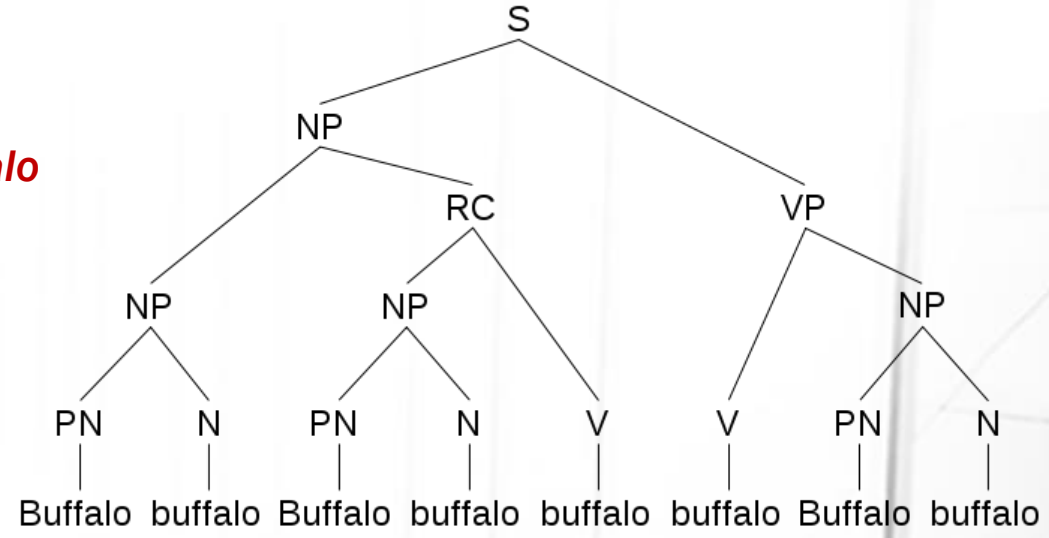
Parsing & Ambiguity

- 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:

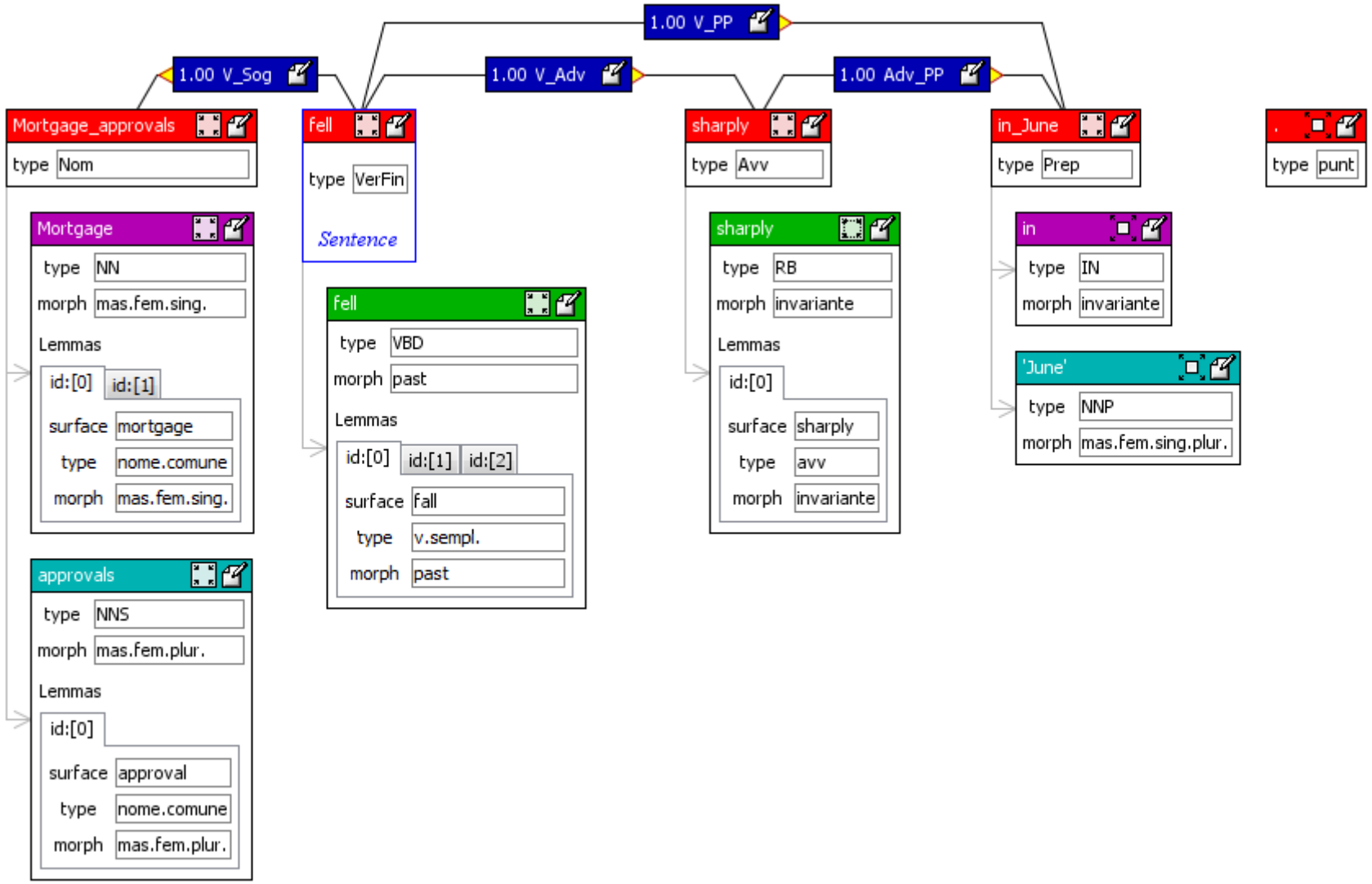
- *To operate in a market* viz. *To operate a body part*

- *Operare in un mercato* ≠ *Operare un paziente*

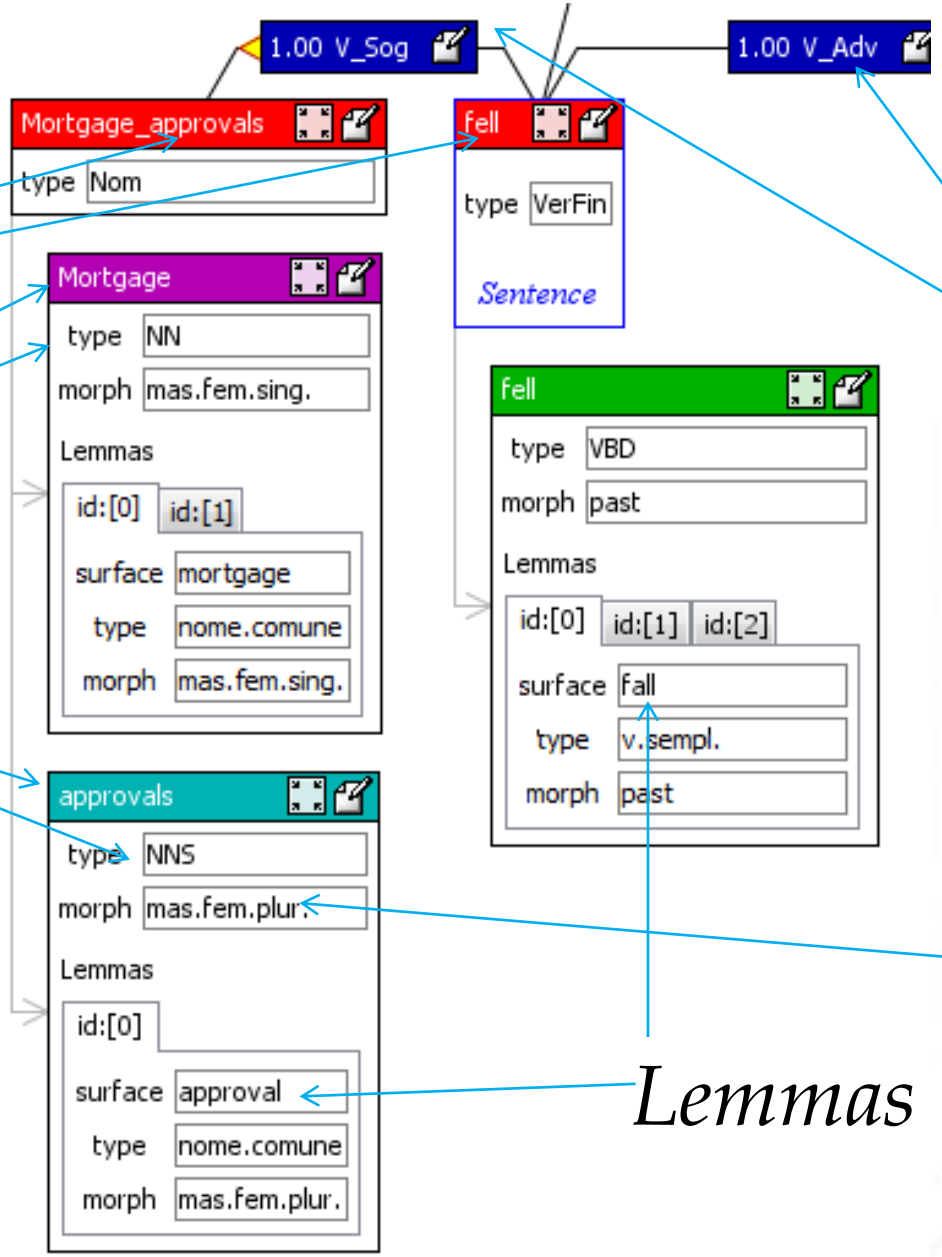


Bison from Buffalo, New York who are intimidated by other bison in their community also happen to intimidate other bison in their community


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



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



Chunks

Tokens and POS tags

Grammatical Relations

Morphological Features

Lemmas

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

Syntax: the role in modern NLP

- Huge influence in the development of the notion of language modeling since 90's through the notion of probabilistic grammars
- Most Statistical Parsing task of 90's, such as sequence labeling via Hiddem Markov Models are still essential, in the application of recurrent neural networks or transformers to texts in NLP taks
- Attention is inspired by the role of the left-right context in grammars
- Positional embeddings are effort to capture position independent grammatical properties
- Word Pieces inherits from the role of «unknown words» in probabilistic grammars

Semantics

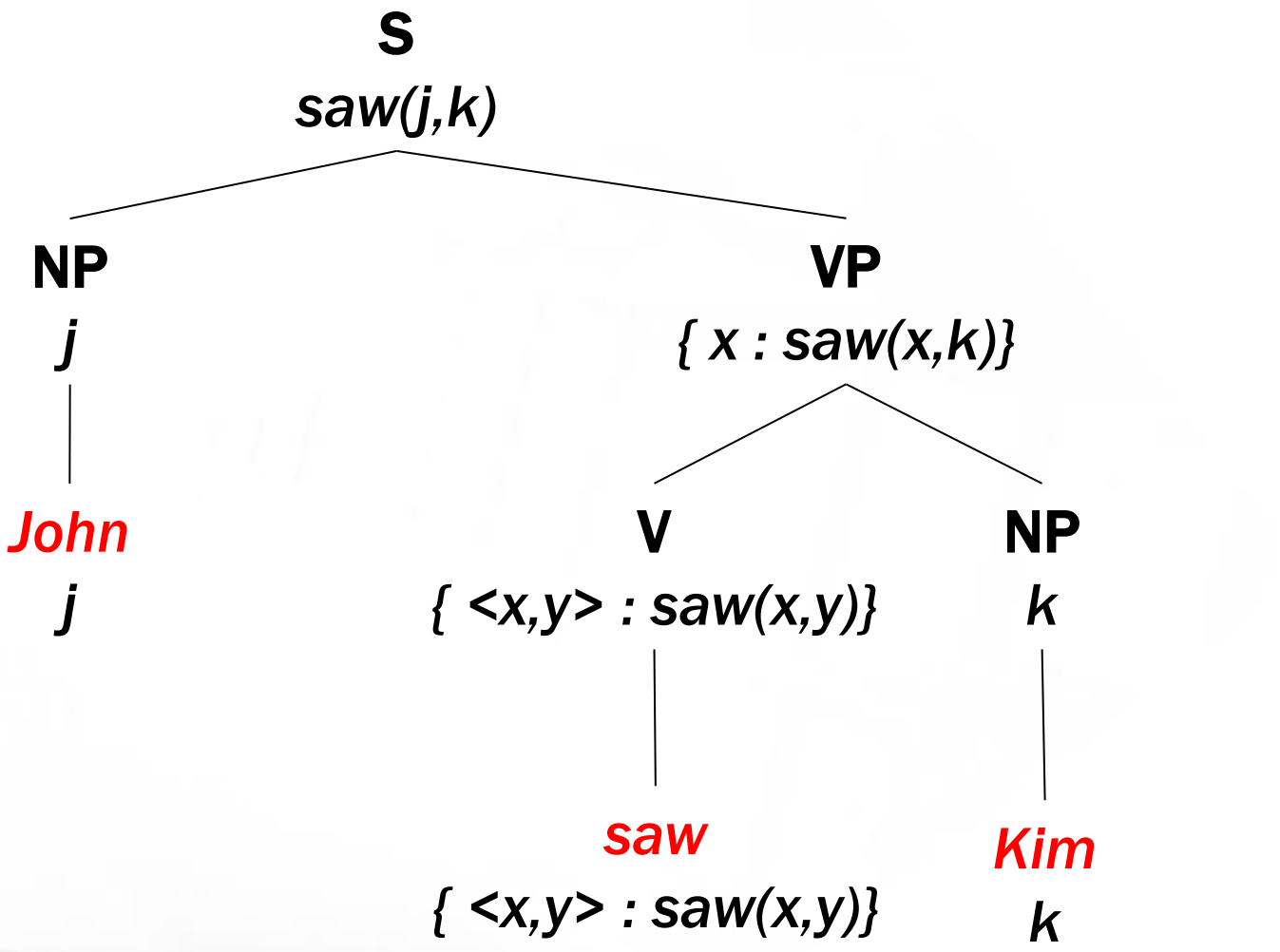
- What is the meaning of the sentence

  John saw Kim?

- Desirable Properties:
  - It should be derivable as a function of the indivdual constituents, i.e. the meanings of costituents such as Kim, John and see
  - Independent from syntactic phenomena, e.g. ***Kim was seen by John*** is a paraphrasis
  - It must be directy 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.

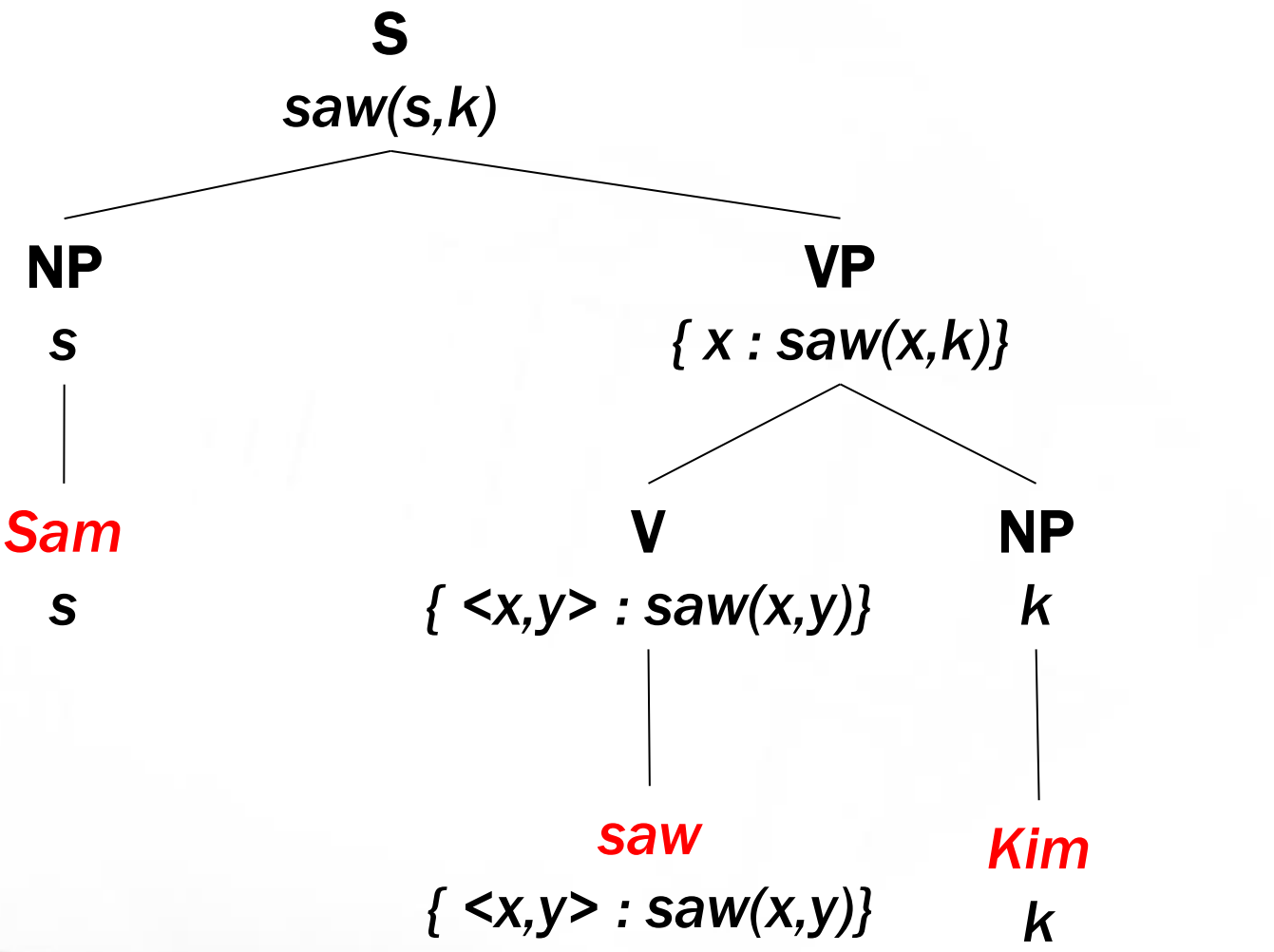
- *"I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages are languages in this sense. Similarly, the set of "sentences" of some formalized system of mathematics can be considered a language"* 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*
  - *Theorem Proving*
  - *SQL queryng from NL expression*
  - *Augmented Geatures for Text Classificatiom, QA and Dialogue*

A Truth conditional semantics

Sam saw Kim



Towards Lambda-calculus

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

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

- *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 neeeded to compute a function f are explicit

- β-reduction: $(\lambda x.\Omega)a \;\; [\Omega]\{x = a\}$
- 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 ia 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 applca-tion 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-Espression	Result
β-reduction:	$(\lambda x.\Omega)a$	$[\Omega]\{x = a\}$
	$(\lambda x.\lambda y.\Omega)(a)(b)$	$\lambda y.\Omega\{x = a\}(b) = [\Omega]\{x = a, y = b\}$

Example: transitive verbs

β-reduction and Compositional Semantics

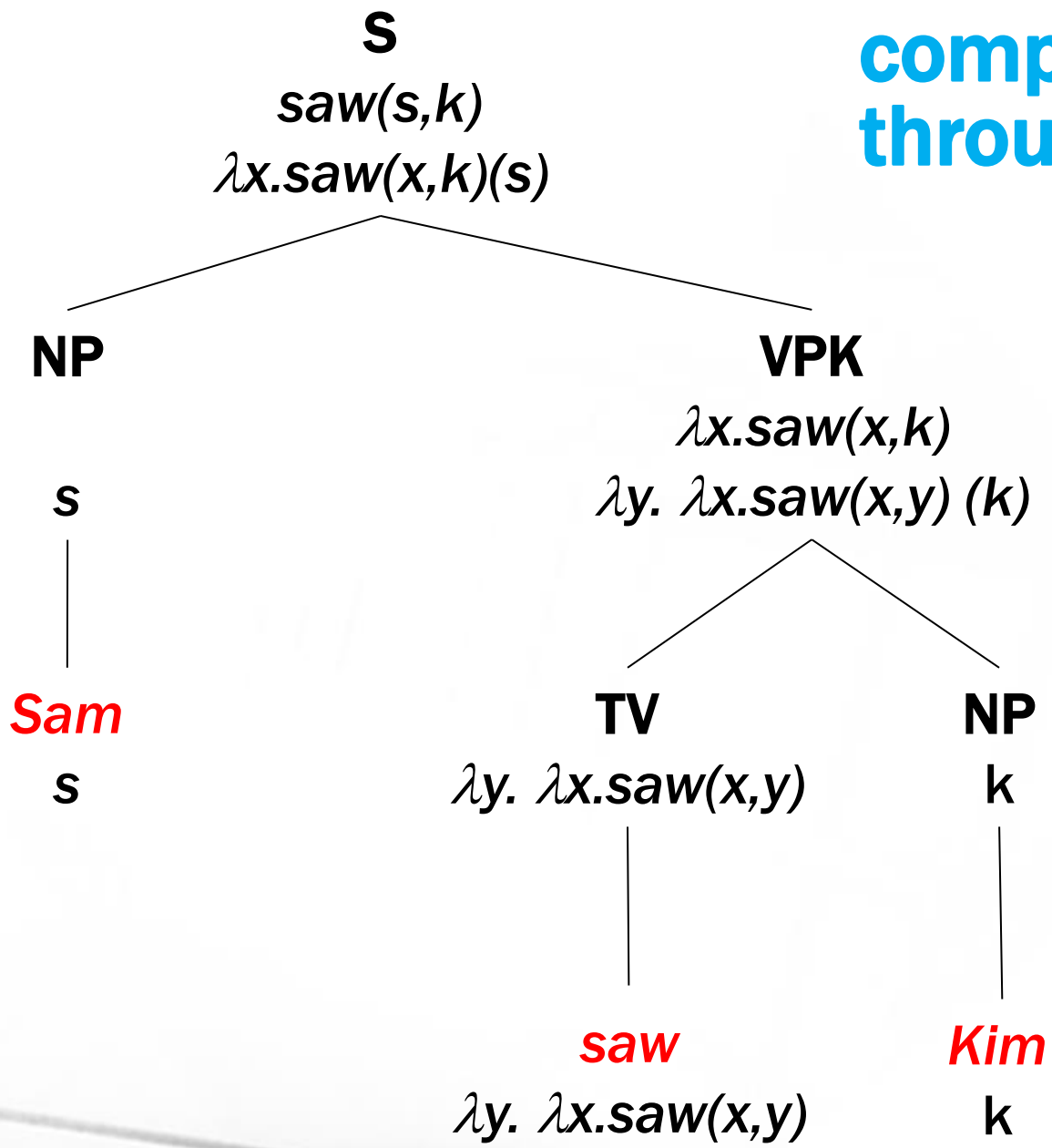
- *Giuseppe runs*: $run(giuseppe)$
- $S \quad \rightarrow \quad NP\ VP$
- Sem Rule1 (*Intransitive verbs*):
  IF The Logical Form (LF) *of NP is NP' and the* LF *of VP is VP'* :
  THEN the LF *S' corresponds to VP'(NP')*
- Consequences:
  runs: $\lambda x.run(x)$
  Giuseppe: $giuseppe$
- $S' = VP'(NP')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

β-reduction and Compositional Semantics (2)

Example: transitive verbs

- *Giuseppe knows Prolog*: $know(giuseppe, prolog)$
- $VP \quad \rightarrow \quad V \ NP$
- 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 espressions*



Sam watched Kim

Sam was seeing Kim

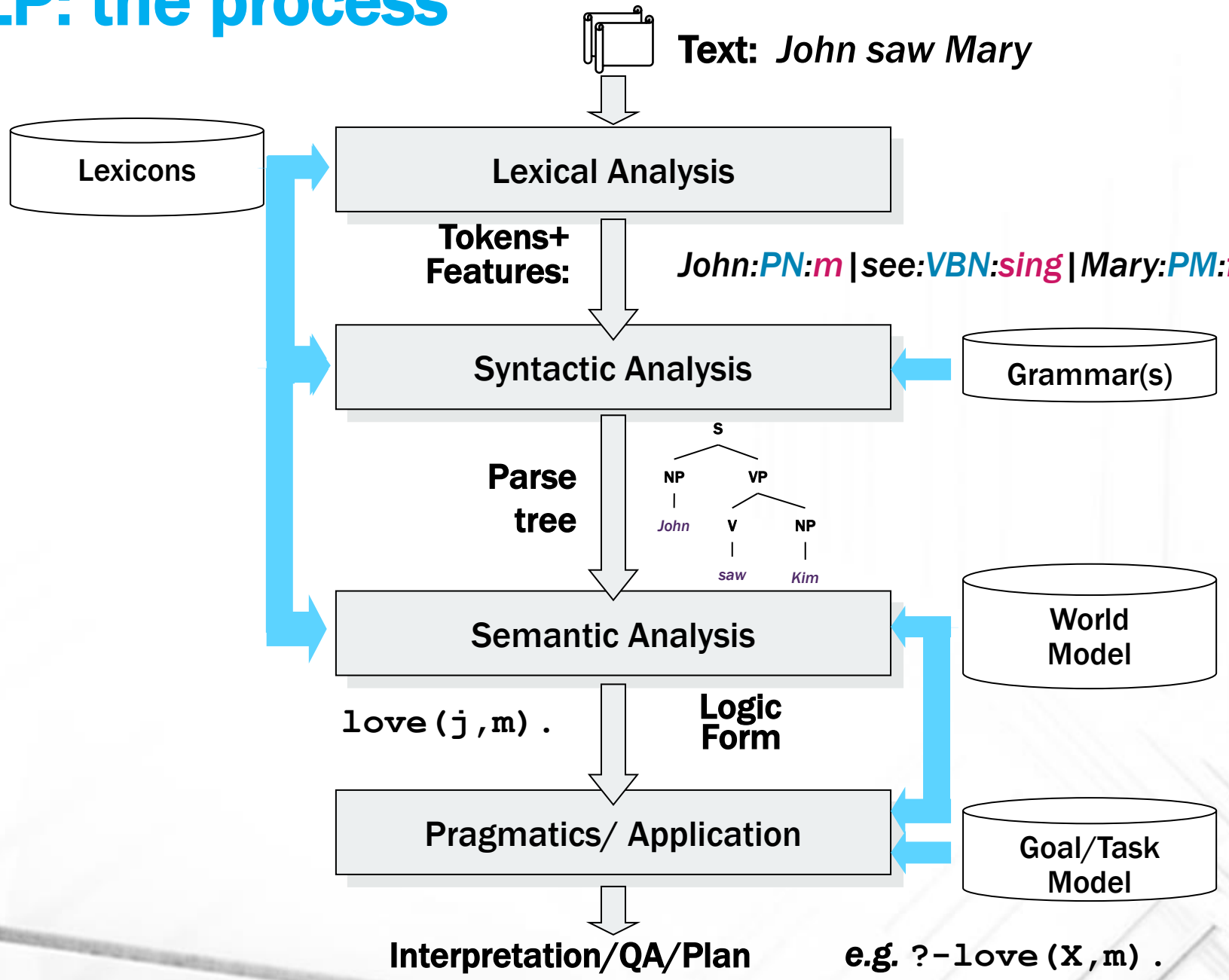
Kim was seen by Sam

John's son watched Kim

Sam saw Jane's daughter

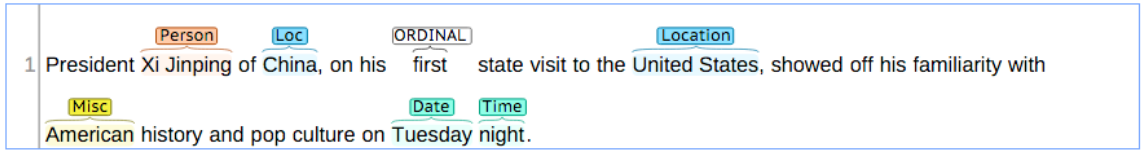
```
saw(s,k)

?-saw(X,k).

X=s
```

NLP: the process

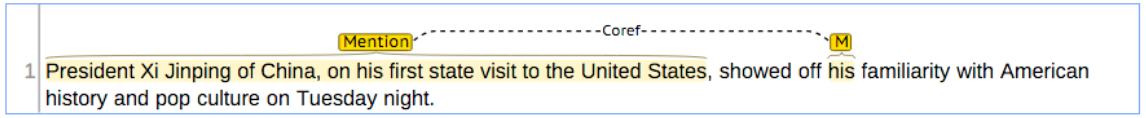


Beyond Parsing:
Named Entity Recognition & Coreference

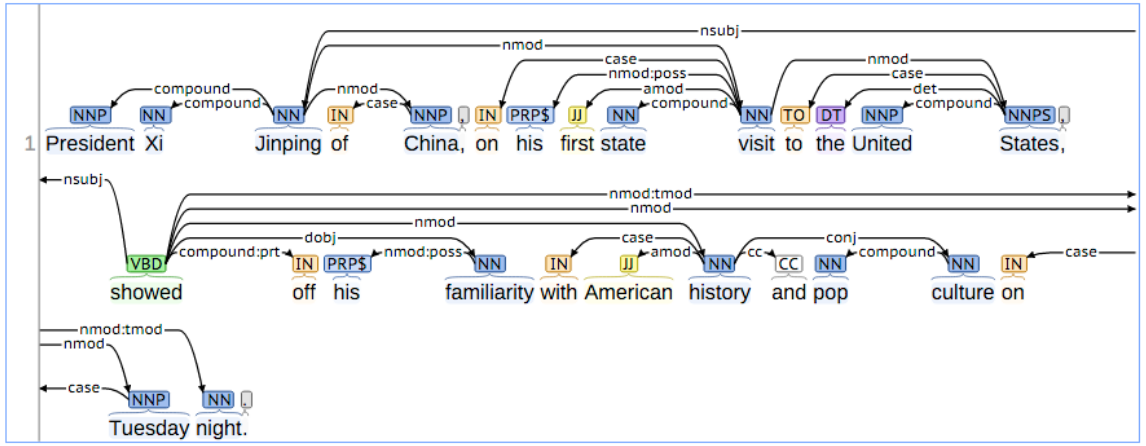
Named Entity Recognition:



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
- Hypernomy
- Hyponomy
- Meronomy

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 bric*
- *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 / H20*
- 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 *H20*
  - I would not say
  - *I like fresh H20 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

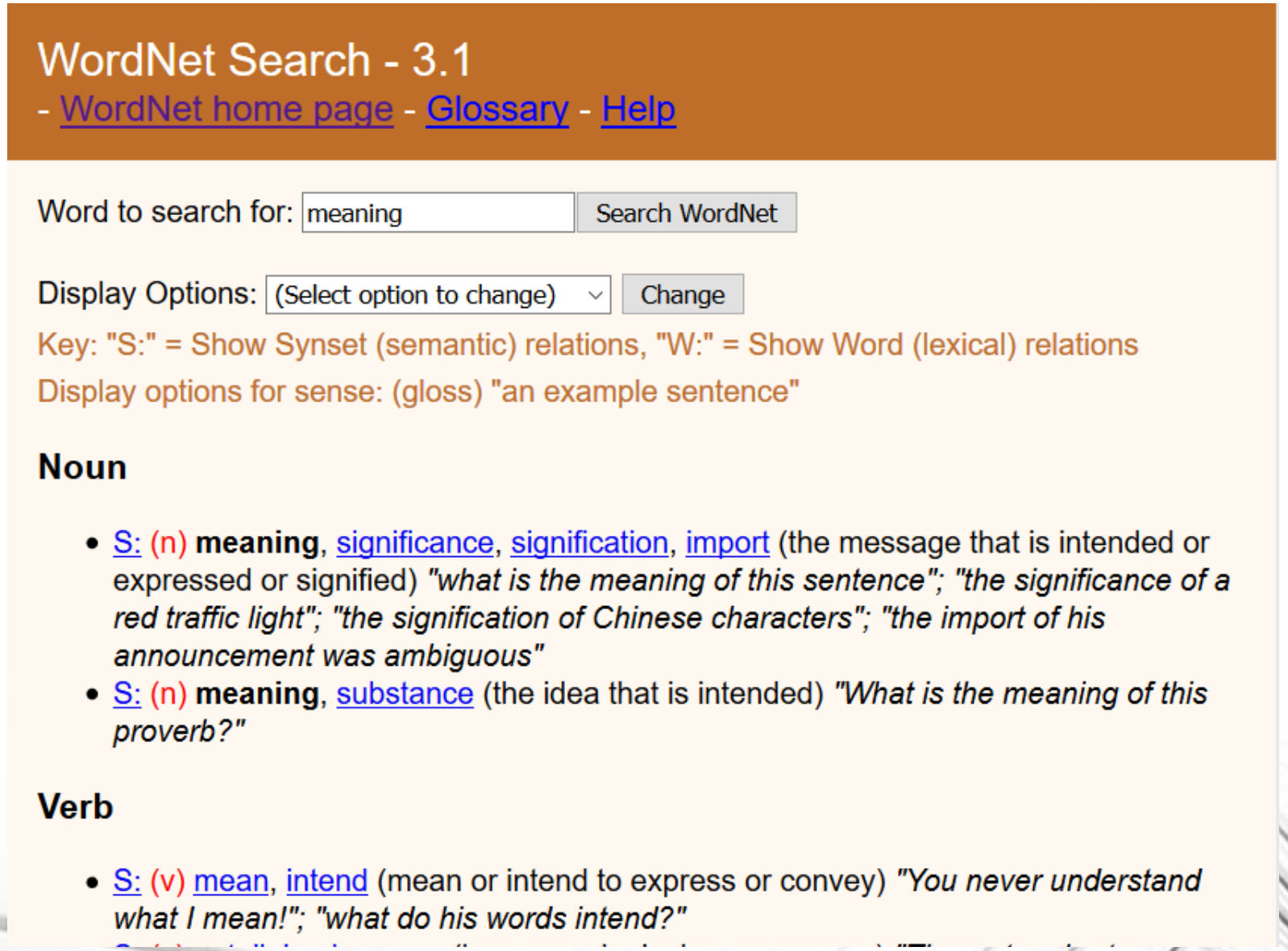
II. WordNet (Miller, 1991)

- 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: meaning Search WordNet

Display Options: (Select option to change) Change

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

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

Noun

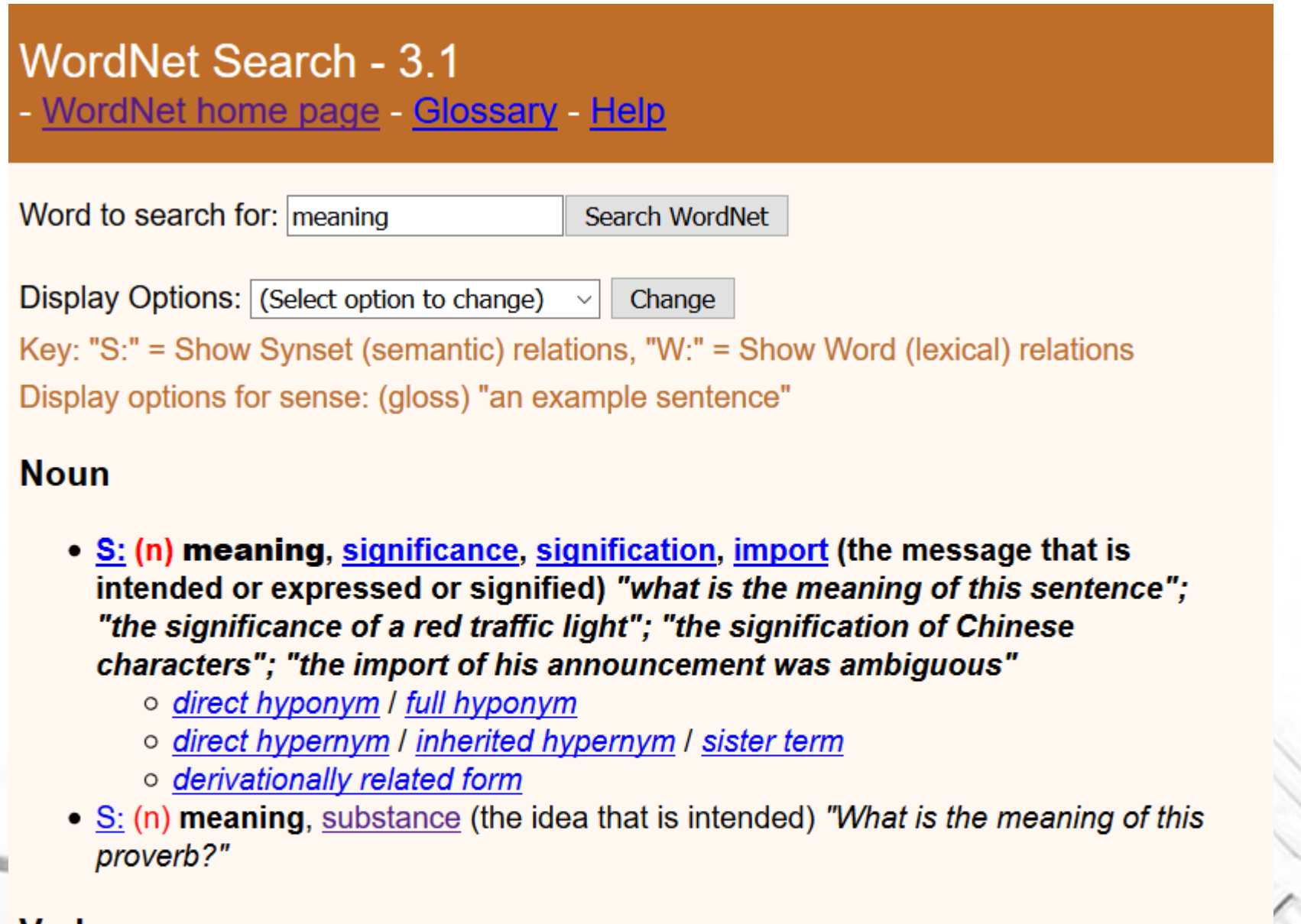
- S: (n) **meaning**, significance, signification, import (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
- S: (n) **meaning**, 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: hyponyms of the word sense *meaning*$_1$

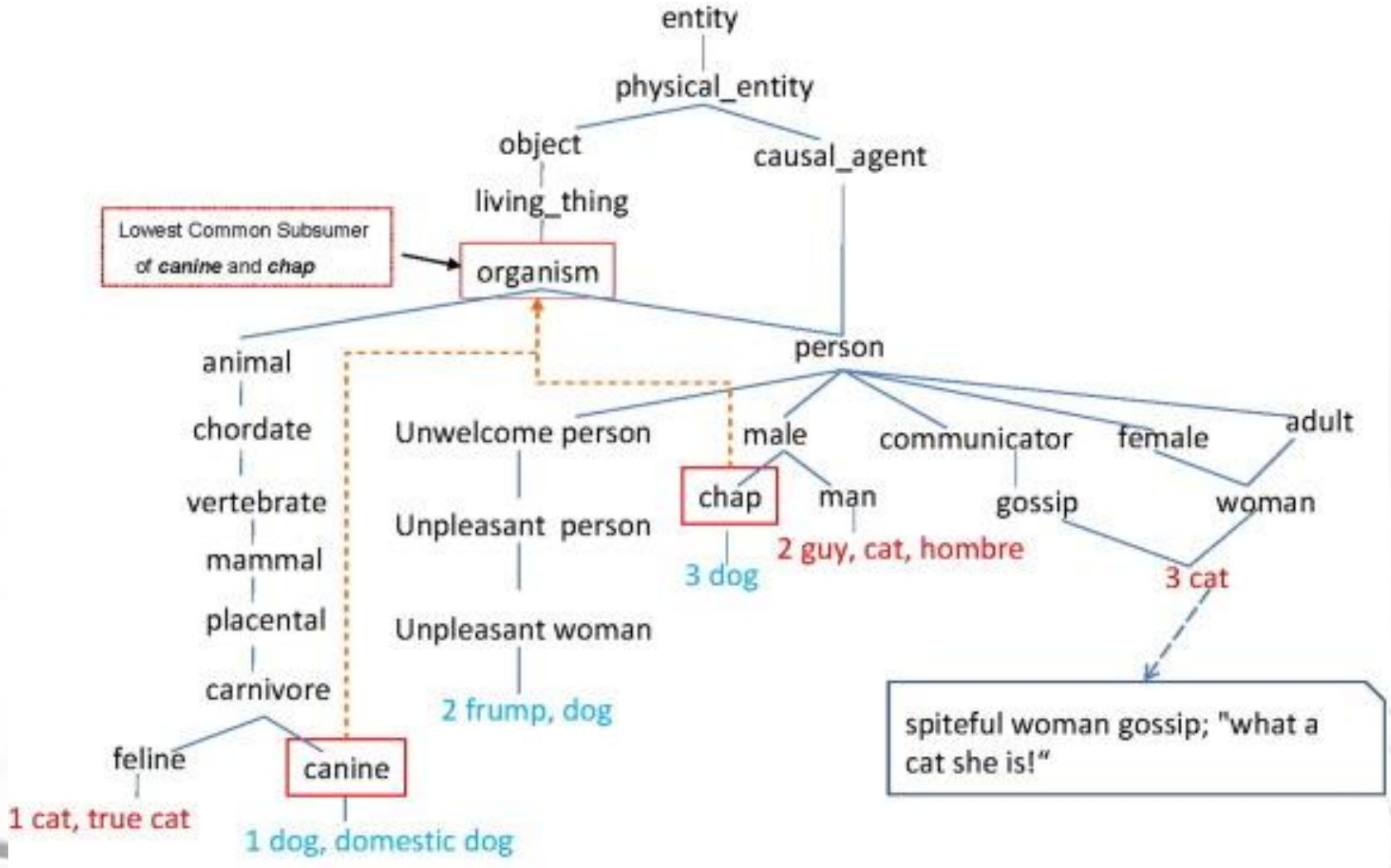
- S: (n) **meaning**, significance, signification, import (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
  - ***direct 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

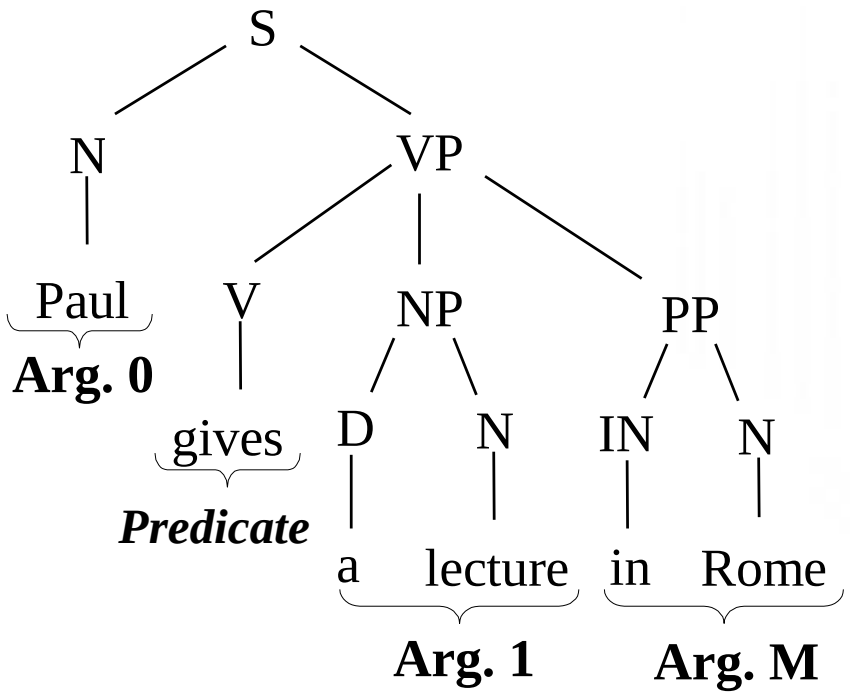


Formal, or sentential, Semantics

- 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 manteinance 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

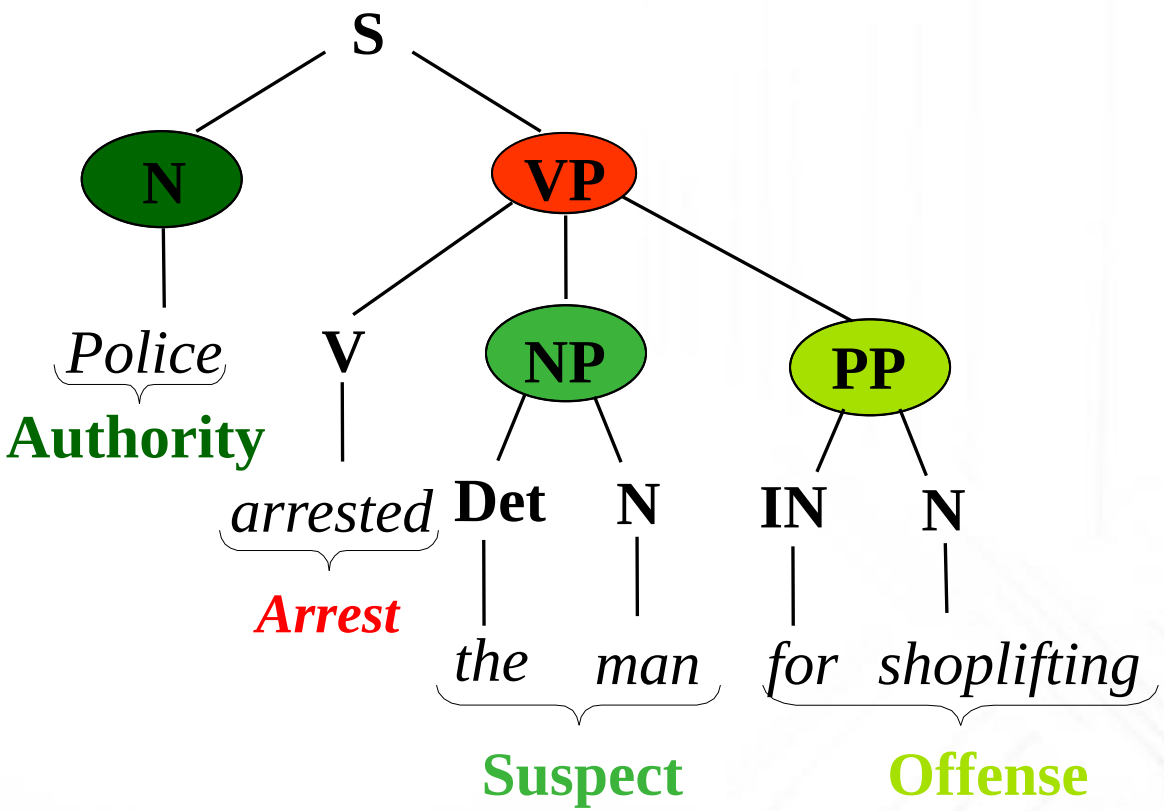
- The syntax-semantic *mapping*
- *Are there any general formalism to denote predicates?*



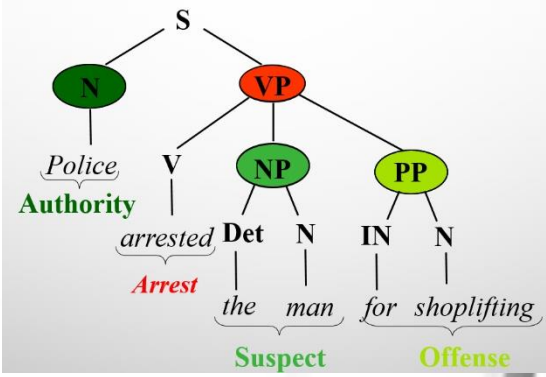
Different formalisms (Annotation schemes) are used as a reference model for predicates: PropBank vs. FrameNet

Linking syntax to semantics: the Framenet style

- *Police arrested the man for shoplifting*



Framenet labeling: the tabular vision



Word	Predicate	Semantic Role	Semantic Role
Police	-	AUTHORITY	-
arrested	Target	**Arrest**	-
the	-	SUSPECT	
man	-	SUSPECT	
for	-	OFFENSE	
shoplifting	Target	OFFENSE	THEFT

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** drowned Martha.
	VICTIM	John drowned **Martha**.
	MEANS	The flood exterminated the rats **by cutting off access to food.**
	CAUSE	**The rockslide** killed nearly half of the climbers.
	INSTRUMENT	It's difficult to suicide **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, . . .	

Lexical Units

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 (SVM-HMM based approach from SAG, (Croce et al, 2012))

- Example

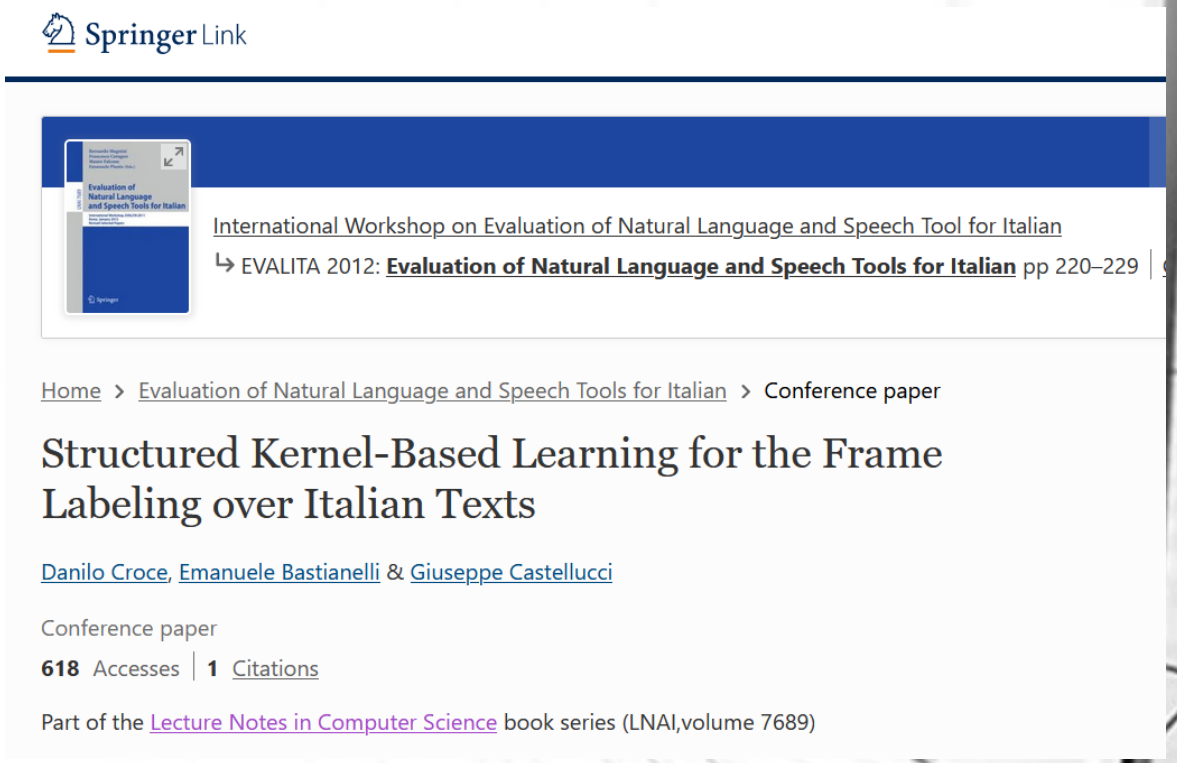
A law enforcement official told CNN that the FBI was investigating.

- 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]$_{Message}$* .

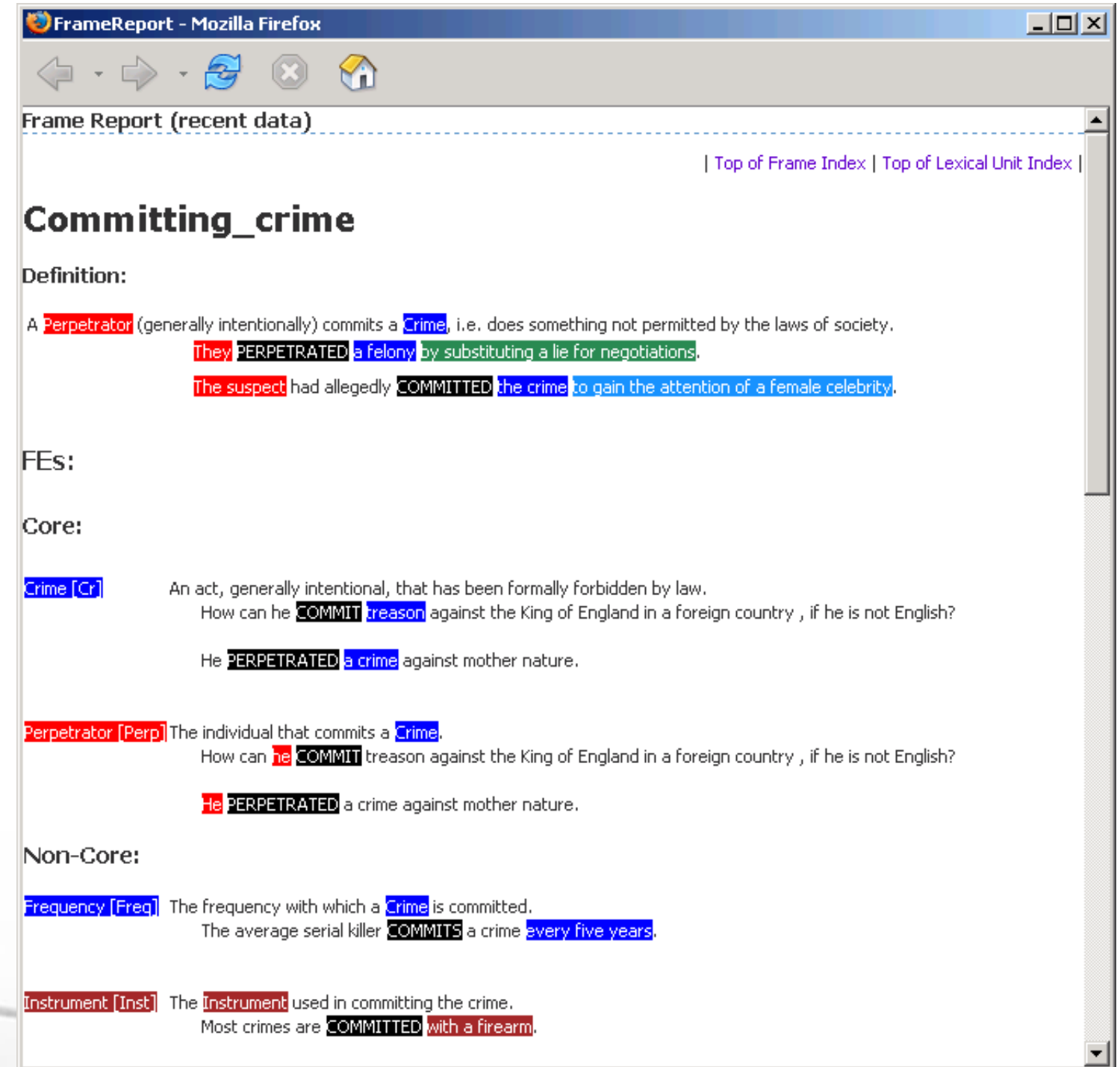
Show CONLL format

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


FrameReport - Mozilla Firefox
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 COMMITS a crime every five years.
Instrument [Inst] The Instrument used in committing the crime.
Most crimes are COMMITTED with a firearm.

Killing

FEs:

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.
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.

Semantic Type: Physical_entity
Excludes: Cause

It's difficult to SUICIDE with only a pocketknife.

Killer [Kill]	The person or sentient entity that causes the death of the Victim.
Means []	The method or action that the Killer or Cause performs resulting in the death of the Victim.
Victim []	The living entity that dies as a result of the killing.

Excludes: Cause

Semantic Type: State_of_affairs
Excludes: Cause

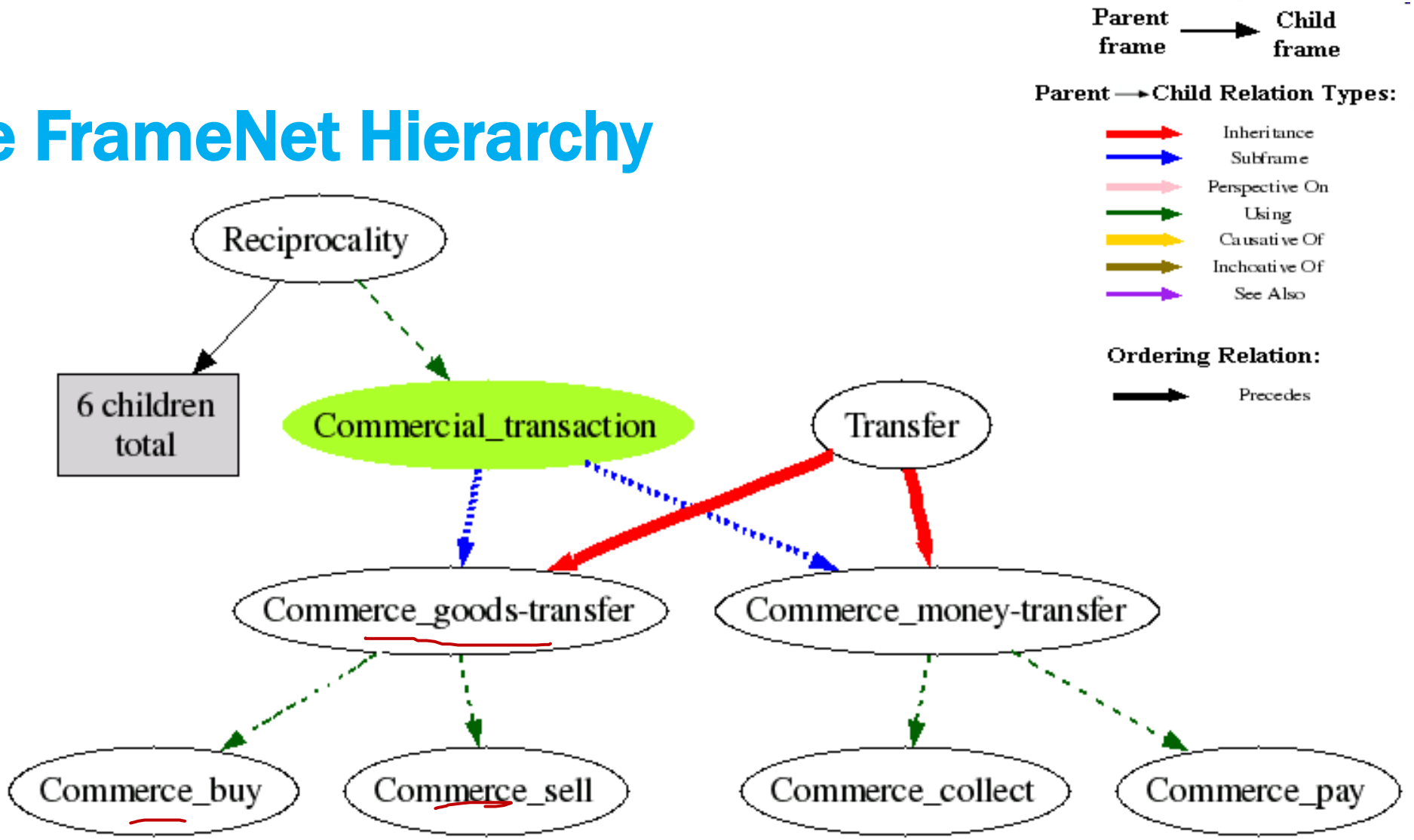
The flood EXTERMINATED the rats by cutting off access to food.

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.

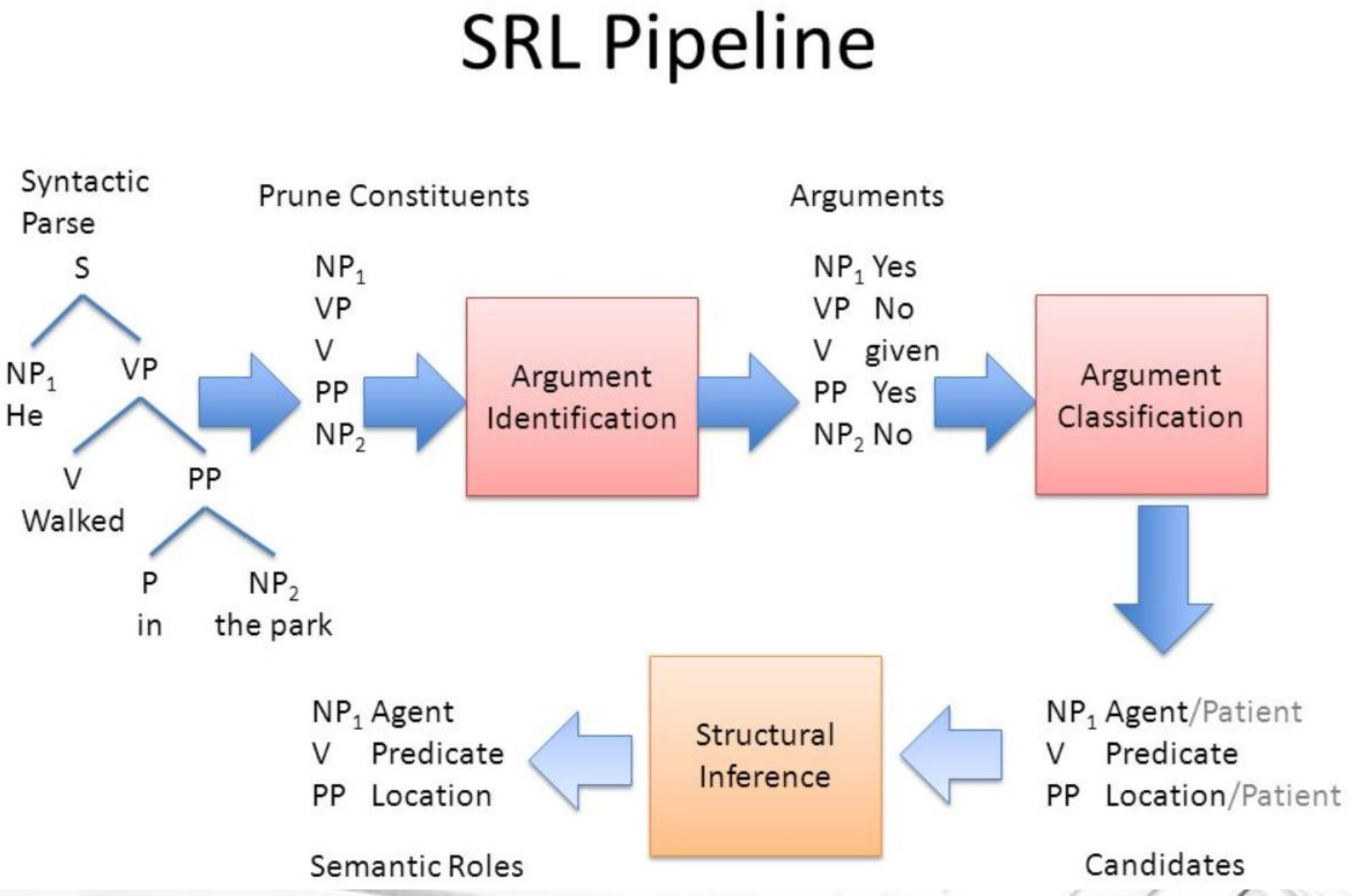
The FrameNet Hierarchy



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
  - >**13,687 lexical units**, >**1,075 frames**, >**174,532 annotated sentences**
- http://framenet.icsi.berkeley.edu

Machine Learning over Framenet/PropBank

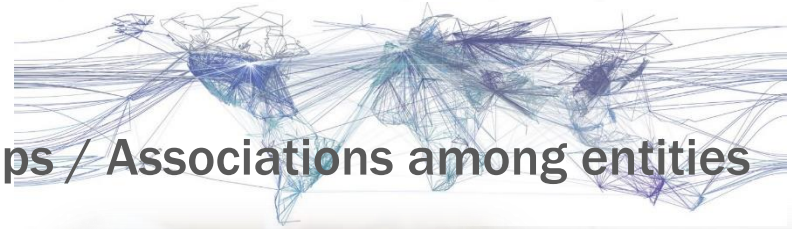


Frame Semantics

- 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

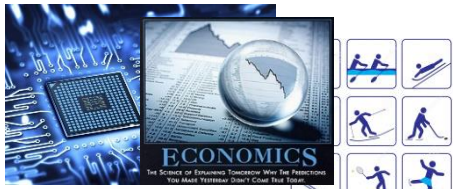
Applications: Target Semantic Phenomena

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

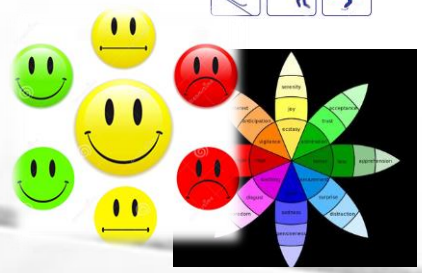


- **Relations.** Relationships / Associations among entities

- **Facts.** Facts and Events

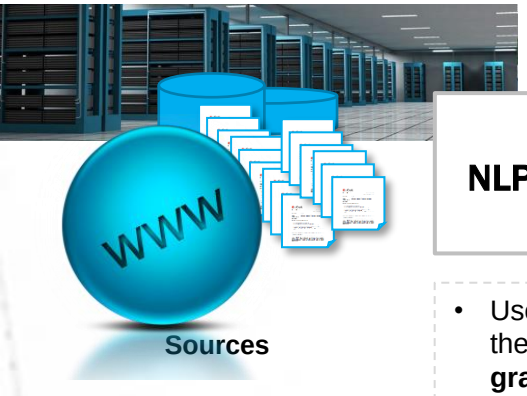


- **Topics.** Discussion topics / Context / Domain



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

NLP Applications: a roadmap



Sources

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**
- **Empricial 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. reccommending) 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 & Schutze, MIT Press 2001.
- Sitografia:
  - SAG, Univ. Roma Tor Vergata: http://sag.art.uniroma2.it/
  - Reveal s.r.l.: http://www.revealsrl.it/