Natural Language Understanding (NLU)

Nikitas N. Karanikolas
Dept. of Informatics and Computer Engineering
University of West Attica, Athens, Greece
nnk@uniwa.gr
http://users.uniwa.gr/nnk

Natural Language Understanding - Tasks

- Phonology
- Morphology
- Syntax
- Semantics
- Representation
- Disambiguation
- Anaphora resolution
- Pragmatics and Discourse analysis
- Resources
- Tools

Phonology

- Phonology is a branch of linguistics concerned with the systematic organization of sounds in languages. It has traditionally focused largely on the study of the systems of phonemes in particular languages
- A phoneme (/ˈfoʊniːm/) is one of the units of sound (or gesture in the case of sign languages, see chereme) that distinguish one word from another in a particular language.
- For example, in most dialects of English, the sound patterns /θʌm/ (*thumb*) and /dʌm/ (*dumb*) are two separate words distinguished by the substitution of one phoneme, /θ/, for another phoneme, /d/.

Phonetic Alphabets

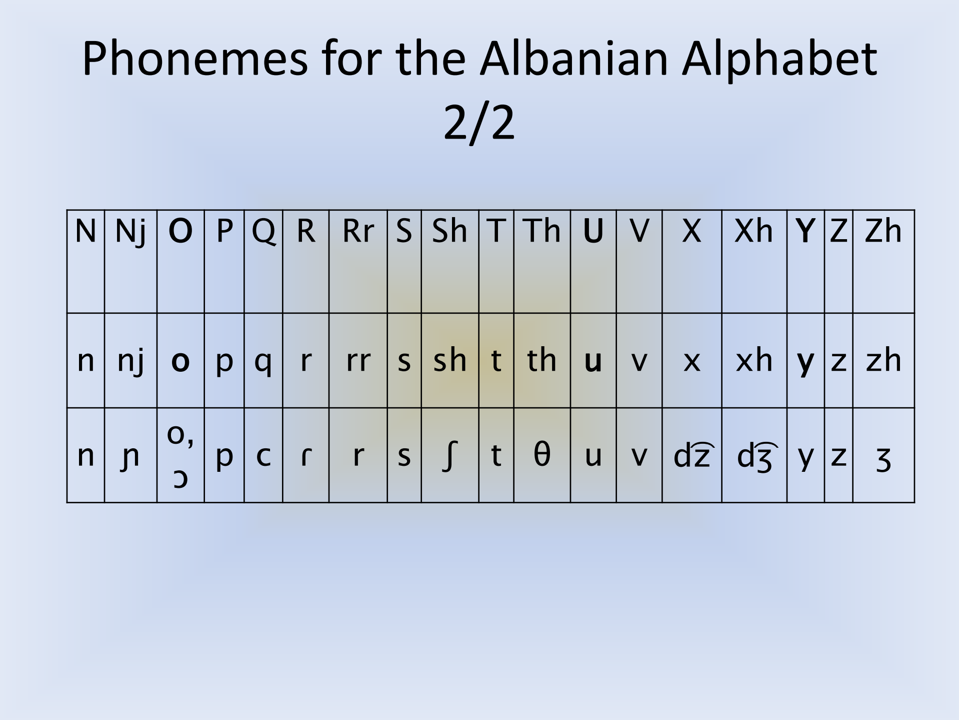
- Why we need phonetic alphabets?
  - To be able to represent graphically all the phonemes exists in every human language
  - To be able to represent with the same symbol a single phoneme that is represented with different letters in different languages
  - To solve the restrictions of the written alphabets
    - γέρος (/ʝeros/)
    - γαρίδα (/ɣariða/)
- How many Exists ?
  - 2, IPA [25] and SAMPA [26]

IPA [25]

- The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin alphabet.
- It was devised by the International Phonetic Association in the late 19th century as a standardized representation of the sounds of spoken language.
- The IPA is used by lexicographers, foreign language students and teachers, linguists, speech-language pathologists, singers, actors, constructed language creators and translators

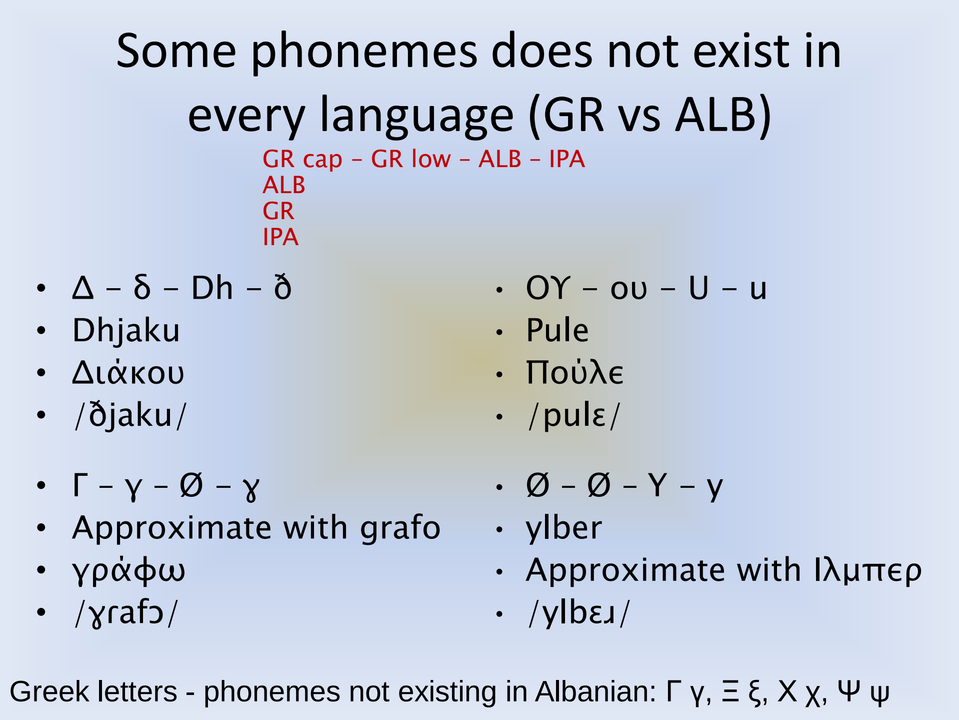
Phonemes for the Albanian Alphabet 1/2

A	B	C	Ç	D	Dh	**E**	**Ë**	F	G	Gj	H	**I**	J	K	L	Ll	M
a	b	c	ç	d	dh	e	ë	f	g	gj	h	i	j	k	l	ll	m
ä	b	t͡s	t͡ʃ	d	ð	e, ɛ	ə, ʌ, ɜ	f	g	ɟ	h	i	j	k	l	ɫ	m



Phonemes for the Albanian Alphabet 2/2

N	Nj	**O**	P	Q	R	Rr	S	Sh	T	Th	**U**	V	X	Xh	**Y**	Z	Zh
n	nj	**o**	p	q	r	rr	s	sh	t	th	**u**	v	x	xh	**y**	z	zh
n	ɲ	o, ɔ	p	c	ɾ	r	s	ʃ	t	θ	u	v	d͡z	d͡ʒ	y	z	ʒ



Some phonemes does not exist in every language (GR vs ALB)

GR cap – GR low – ALB – IPA
ALB
GR
IPA

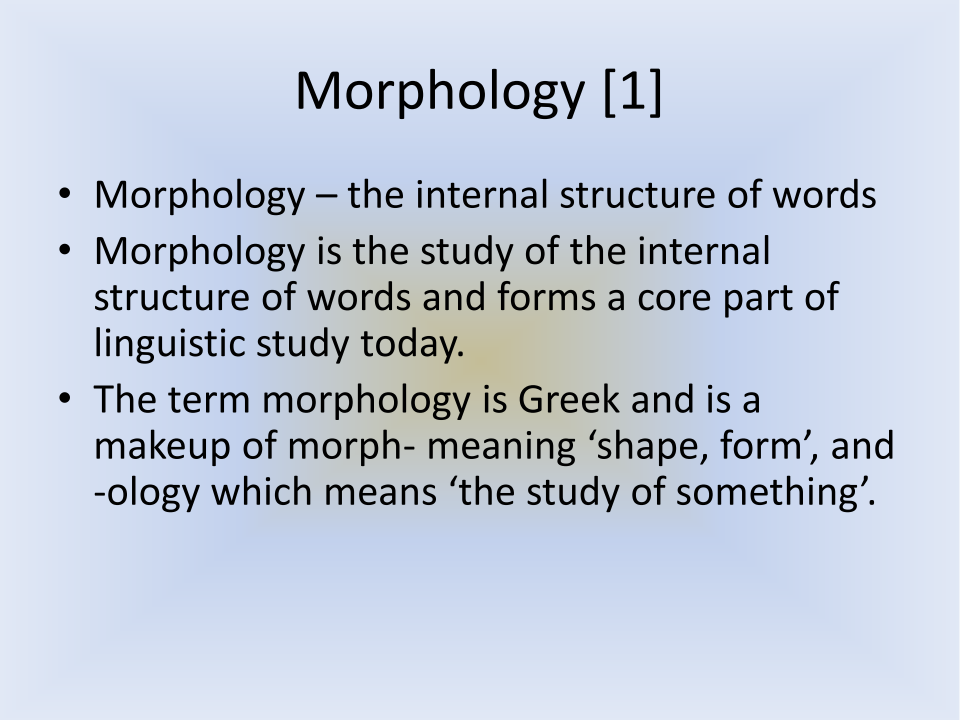
- Δ – δ – Dh – ð
- Dhjaku
- Διάκου
- /ðjaku/

- Γ – γ – Ø – ɣ
- Approximate with grafo
- γράφω
- /ɣrafɔ/

- ΟΥ – ου – U – u
- Pule
- Πούλε
- /pulɛ/

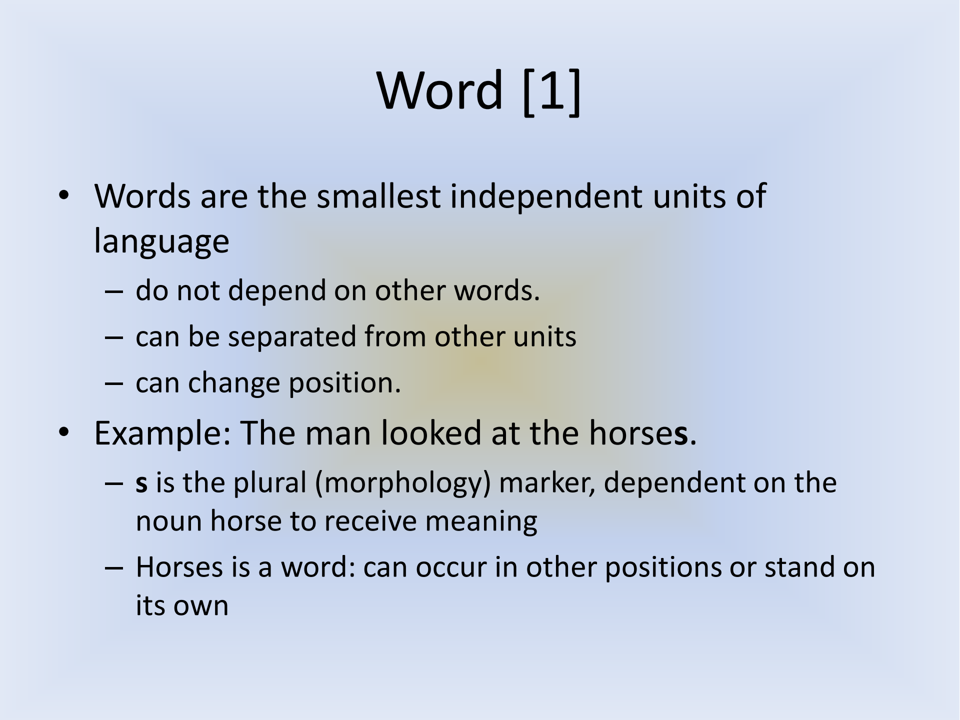
- Ø – Ø – Y – y
- ylber
- Approximate with Ιλμπερ
- /ylbɛɹ/

Greek letters - phonemes not existing in Albanian: Γ γ, Ξ ξ, Χ χ, Ψ ψ



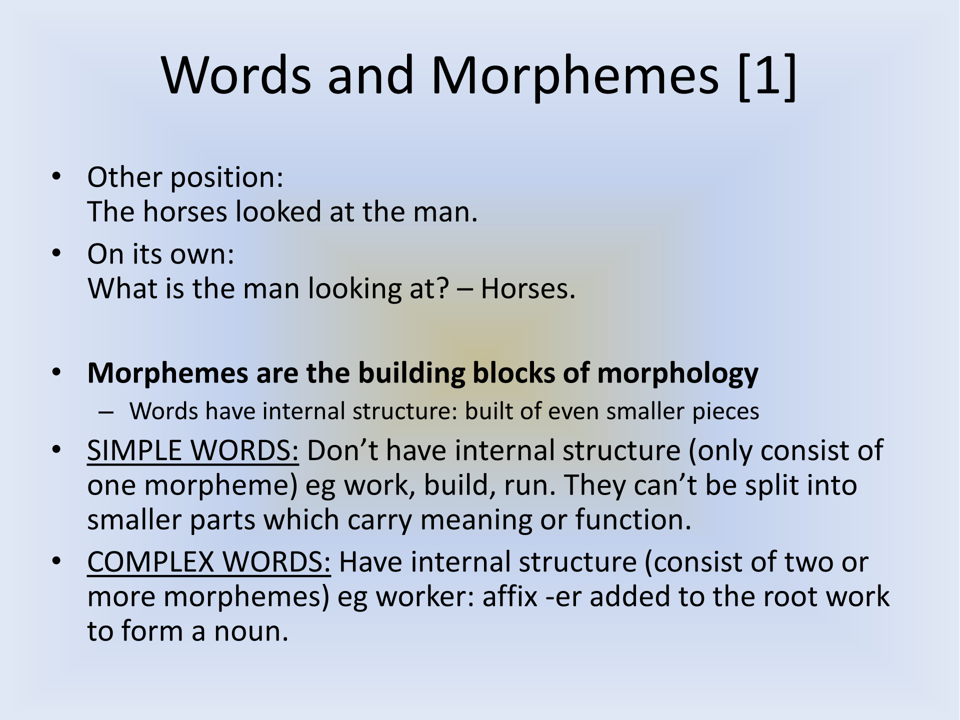
Morphology [1]

- Morphology – the internal structure of words
- Morphology is the study of the internal structure of words and forms a core part of linguistic study today.
- The term morphology is Greek and is a makeup of morph- meaning ‘shape, form’, and -ology which means ‘the study of something’.



Word [1]

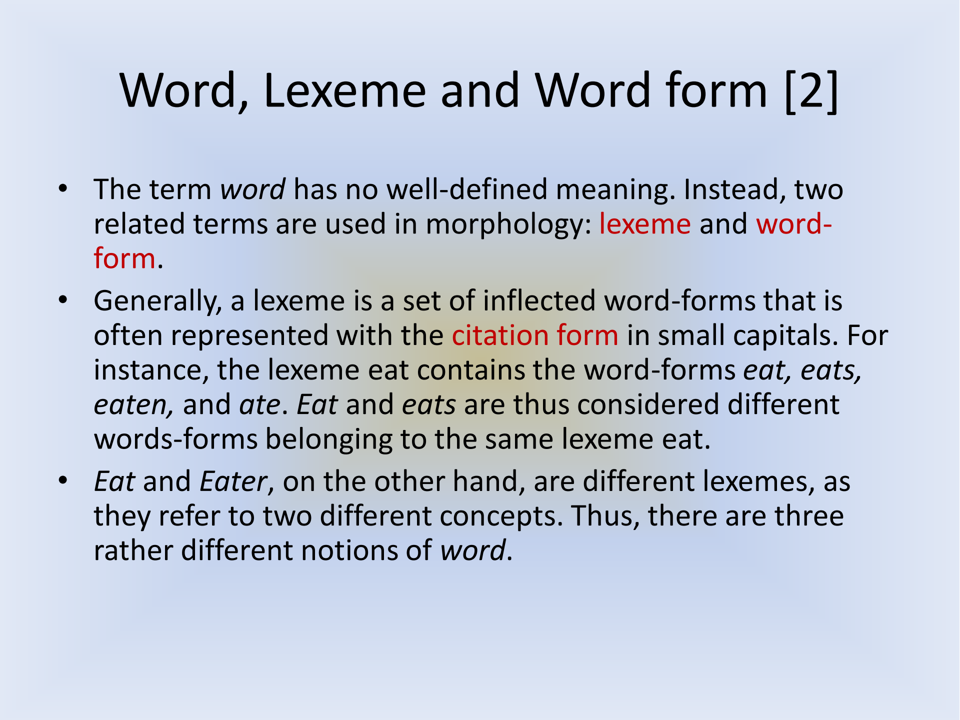
- Words are the smallest independent units of language
  - do not depend on other words.
  - can be separated from other units
  - can change position.
- Example: The man looked at the horse**s**.
  - **s** is the plural (morphology) marker, dependent on the noun horse to receive meaning
  - Horses is a word: can occur in other positions or stand on its own



Words and Morphemes [1]

- Other position:
  The horses looked at the man.
- On its own:
  What is the man looking at? – Horses.

- **Morphemes are the building blocks of morphology**
  - Words have internal structure: built of even smaller pieces
- SIMPLE WORDS: Don't have internal structure (only consist of one morpheme) eg work, build, run. They can't be split into smaller parts which carry meaning or function.
- COMPLEX WORDS: Have internal structure (consist of two or more morphemes) eg worker: affix -er added to the root work to form a noun.



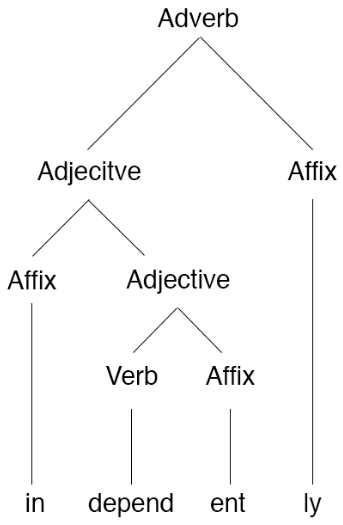
Word, Lexeme and Word form [2]

- The term *word* has no well-defined meaning. Instead, two related terms are used in morphology: lexeme and word-form.
- Generally, a lexeme is a set of inflected word-forms that is often represented with the citation form in small capitals. For instance, the lexeme eat contains the word-forms *eat, eats, eaten,* and *ate*. *Eat* and *eats* are thus considered different words-forms belonging to the same lexeme eat.
- *Eat* and *Eater*, on the other hand, are different lexemes, as they refer to two different concepts. Thus, there are three rather different notions of *word*.

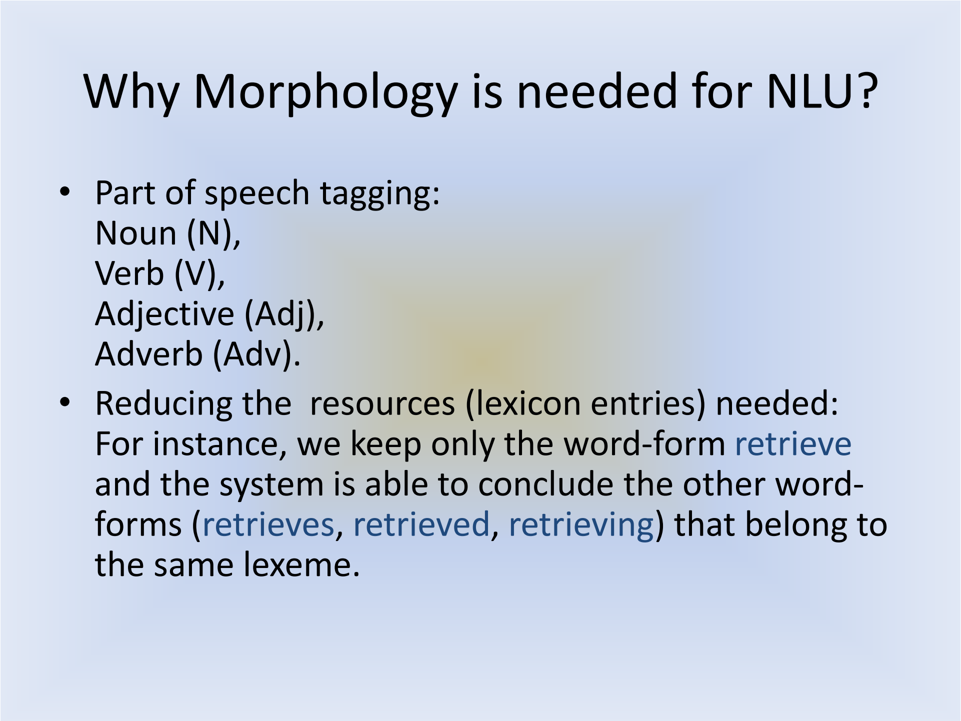
Inflection vs. word formation [2]

- Given the notion of a lexeme, it is possible to distinguish two kinds of morphological rules. Some morphological rules relate to different forms of the same lexeme; while other rules relate to different lexemes.
- Rules of the first kind are inflectional rules, while those of the second kind are rules of word formation.
- The generation of the English plural *dogs* from *dog* is an inflectional rule, while compound phrases and words like *dog catcher* or *dishwasher* are examples of word formation.
- Informally, word formation rules form *new words* (more accurately, *new lexemes*), while inflection rules yield *variant forms of the same word* (same lexeme).

A morphology tree [2]

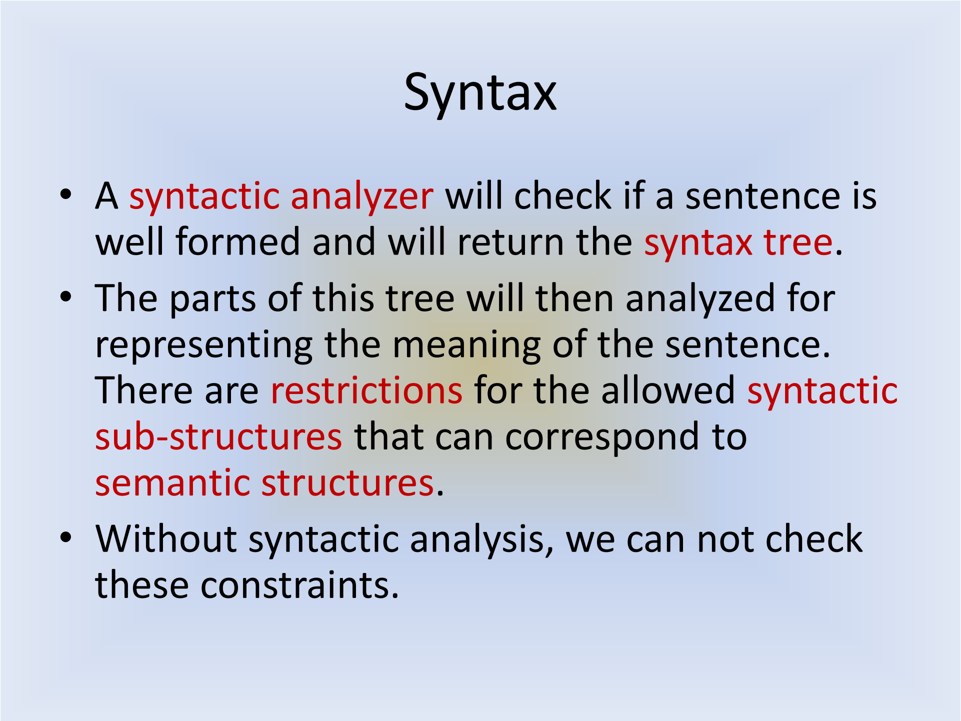


- In, ent and ly are morphemes
- Depend (adj), Independ (adj), Independent (adj) and Independently (adverb) are lexemes



Why Morphology is needed for NLU?

- Part of speech tagging:
  Noun (N),
  Verb (V),
  Adjective (Adj),
  Adverb (Adv).
- Reducing the resources (lexicon entries) needed: For instance, we keep only the word-form retrieve and the system is able to conclude the other word-forms (retrieves, retrieved, retrieving) that belong to the same lexeme.



Syntax

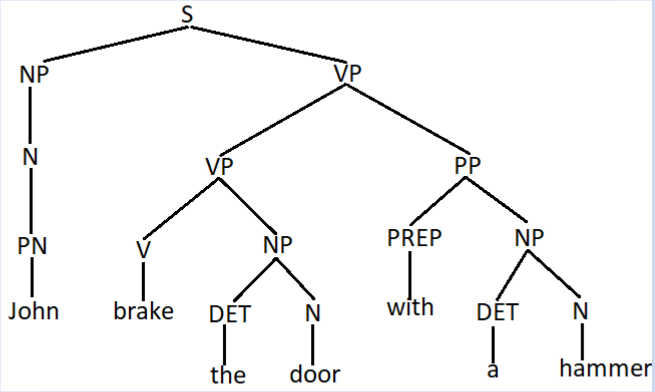
- A syntactic analyzer will check if a sentence is well formed and will return the syntax tree.
- The parts of this tree will then analyzed for representing the meaning of the sentence. There are restrictions for the allowed syntactic sub-structures that can correspond to semantic structures.
- Without syntactic analysis, we can not check these constraints.

A syntax tree

- John broke the door with a hammer

```
s(   np(n(pn(John))),
     vp(    vp(    v(brake,past).
                   np(det(the),n(door)))
            pp(    prep(with),
                   np(det(a), n(hammer)))
     )
)
```

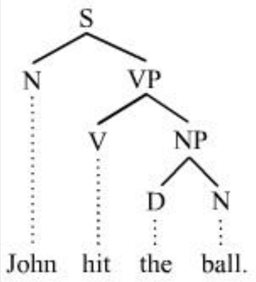
Derivation Tree



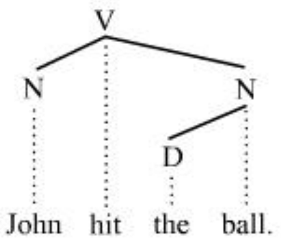
Grammars

- Phrase Structure Grammars and Rules [13]
  - *constituency-based*
    binary division of the clause into subject (noun phrase NP) and predicate (verb phrase VP)
  - a one-to-one-or-more correspondence
  - A → B C; A → (B) C; A→ {B, C}
  - S → NP VP; NP → (Det) N1; N1 → (AP) N1 (PP)
  - Can be context free or context sensitive
- Dependency Grammars [24]
  - dependency relation
    understanding of sentence logic in terms of predicates and their arguments
  - a one-to-one relation

Parse tree VS Dependency tree [12]



Constituency-based parse tree



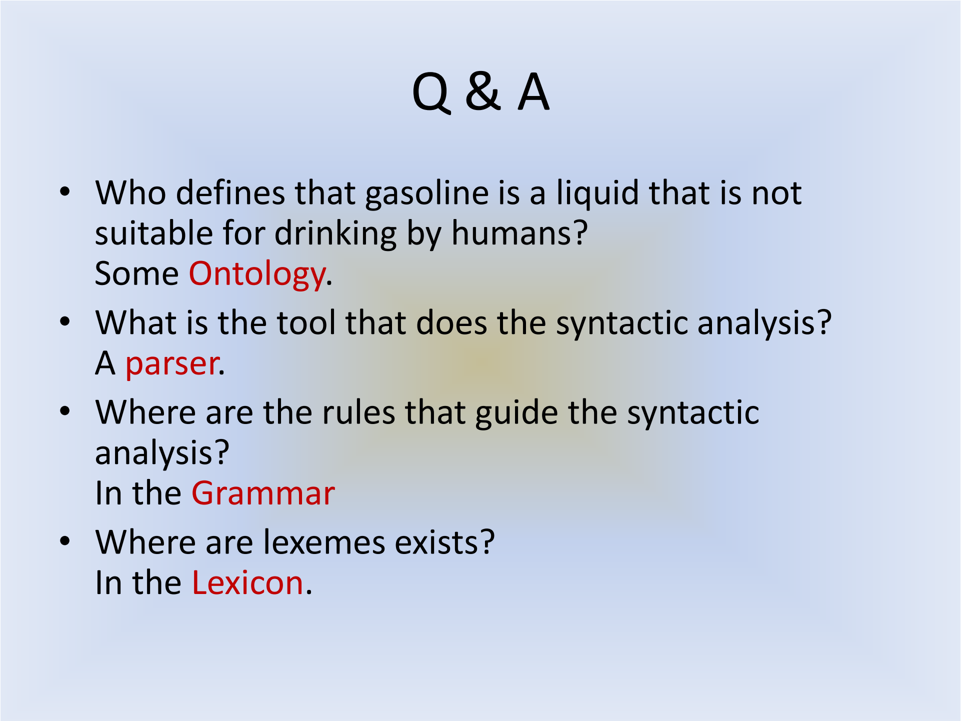
Dependency-based parse tree

More for Grammars and Parsing

- Augmented Transition Networks (ATNs) [16]
- Generalized Phrase Structure Grammar (GPSG) [14]
- Lexical-Function grammar (LFG) [18]
- Head-Driven Phrase Structure Grammar (HDPGs) [16]
- Categorial Grammar (CG) [17]
- Shallow parsing [15]

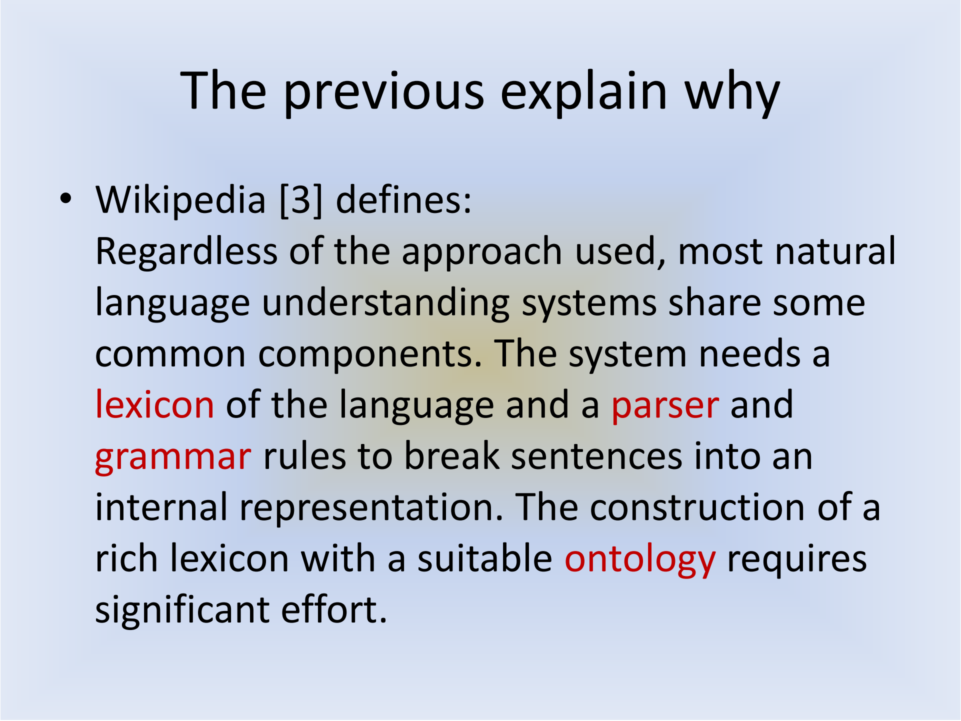
Semantics

- A well formed syntactically sentence is not always correct
- “John drunk 3 liters gasoline” is syntactically correct but gasoline is a liquid that is not suitable for drinking by humans and John is a human.
- The semantics are what recognizes that John is a proper name and consequently it refers to a human and that gasoline is a liquid that is not a kind of food or beverage
- There is also some semantic restriction that the consumed item in some verb of feeding should be food or beverage



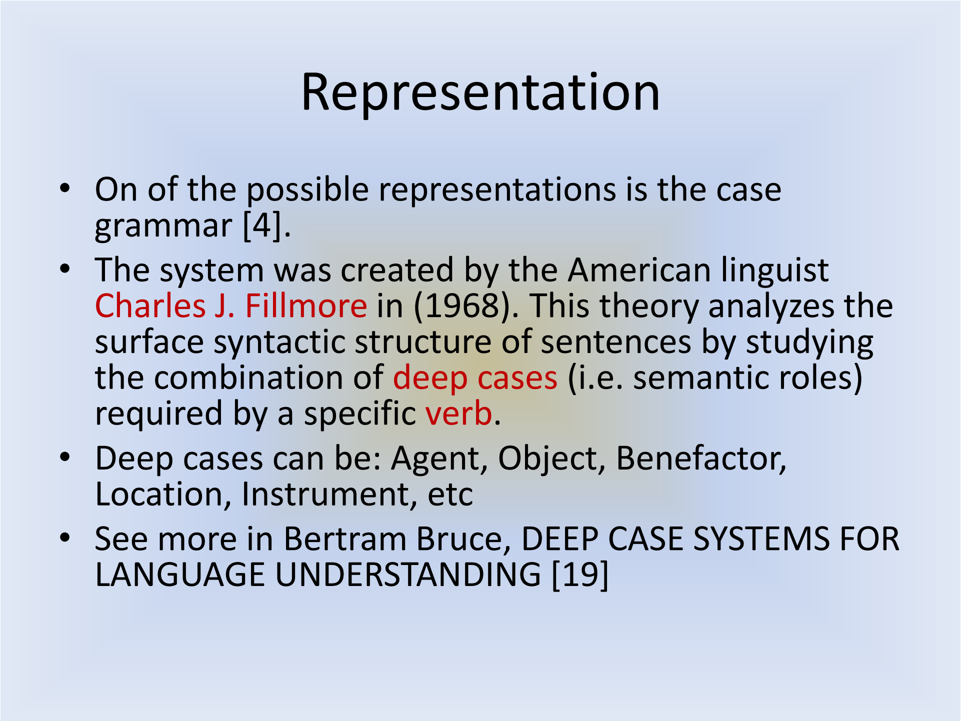
Q & A

- Who defines that gasoline is a liquid that is not suitable for drinking by humans?
  Some Ontology.
- What is the tool that does the syntactic analysis?
  A parser.
- Where are the rules that guide the syntactic analysis?
  In the Grammar
- Where are lexemes exists?
  In the Lexicon.



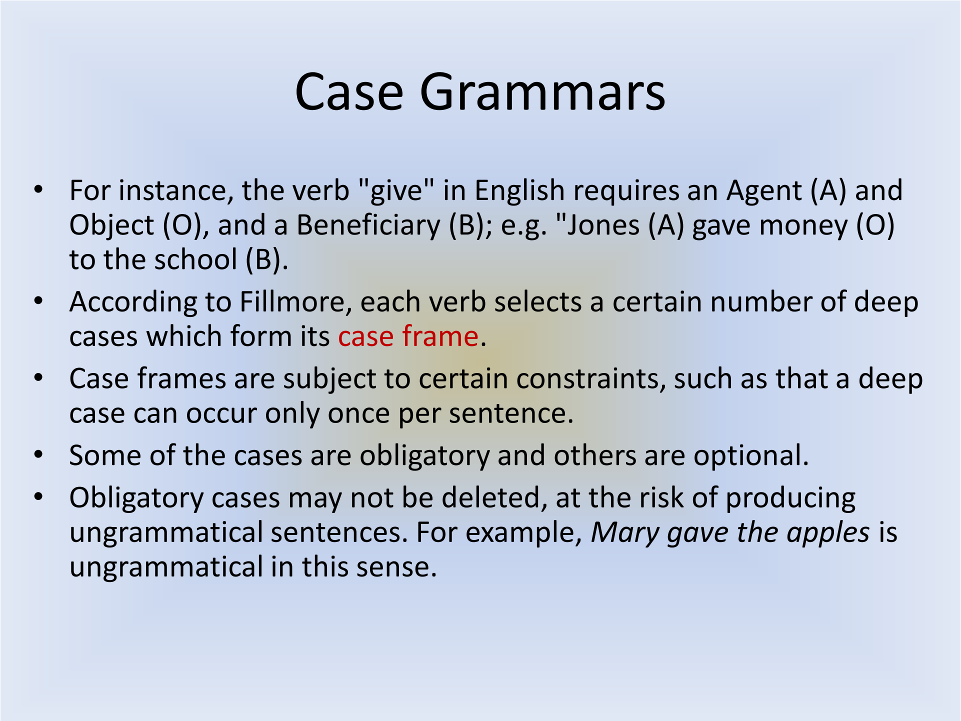
The previous explain why

- Wikipedia [3] defines:
  Regardless of the approach used, most natural language understanding systems share some common components. The system needs a lexicon of the language and a parser and grammar rules to break sentences into an internal representation. The construction of a rich lexicon with a suitable ontology requires significant effort.



Representation

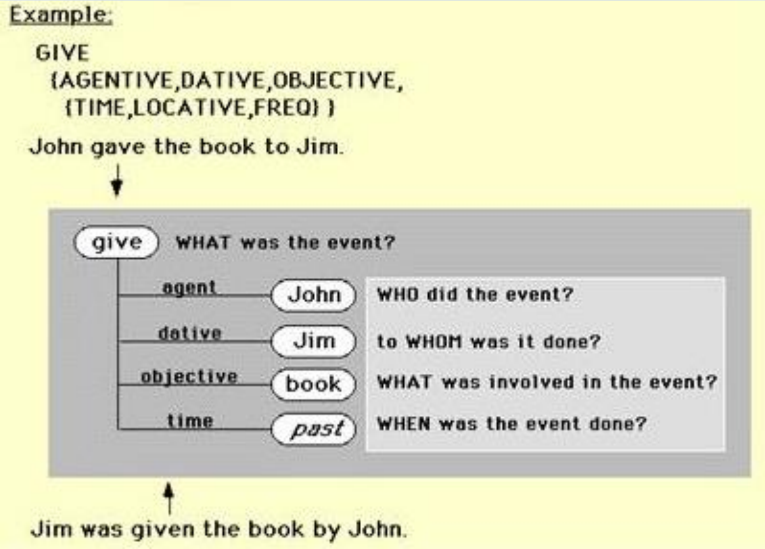
- On of the possible representations is the case grammar [4].
- The system was created by the American linguist Charles J. Fillmore in (1968). This theory analyzes the surface syntactic structure of sentences by studying the combination of deep cases (i.e. semantic roles) required by a specific verb.
- Deep cases can be: Agent, Object, Benefactor, Location, Instrument, etc
- See more in Bertram Bruce, DEEP CASE SYSTEMS FOR LANGUAGE UNDERSTANDING [19]



Case Grammars

- For instance, the verb "give" in English requires an Agent (A) and Object (O), and a Beneficiary (B); e.g. "Jones (A) gave money (O) to the school (B).
- According to Fillmore, each verb selects a certain number of deep cases which form its case frame.
- Case frames are subject to certain constraints, such as that a deep case can occur only once per sentence.
- Some of the cases are obligatory and others are optional.
- Obligatory cases may not be deleted, at the risk of producing ungrammatical sentences. For example, *Mary gave the apples* is ungrammatical in this sense.

Case Grammar Example, from [5]



Example:

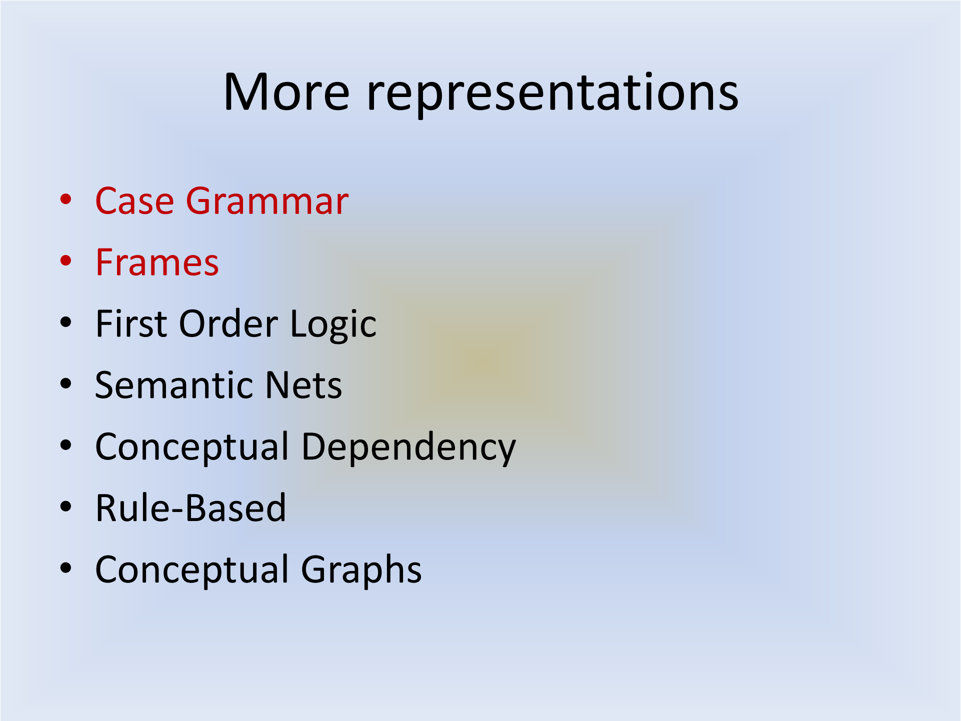
GIVE
{AGENTIVE,DATIVE,OBJECTIVE,
{TIME,LOCATIVE,FREQ} }

John gave the book to Jim.

Jim was given the book by John.

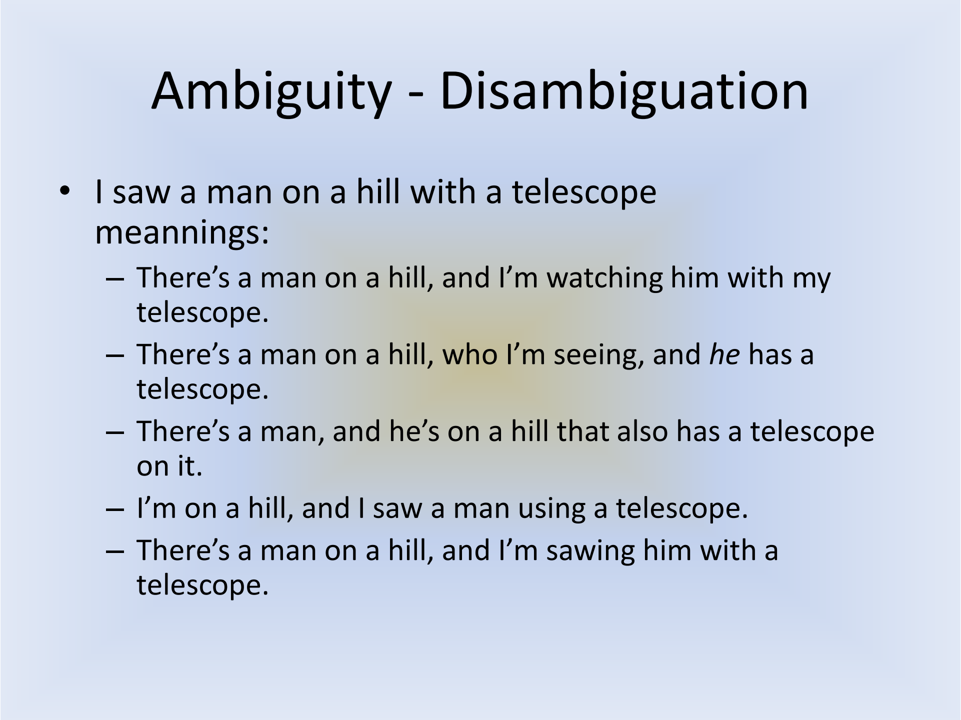
Frames

- A frame language [20] is a technology used for knowledge representation in artificial intelligence. Frames are stored as ontologies of sets and subsets of the frame concepts.
- They are similar to class hierarchies in object-oriented languages although their fundamental design goals are different. Frames are focused on explicit and intuitive representation of knowledge whereas objects focus on encapsulation and information hiding
- Frames originated in AI research and objects primarily in software engineering. However, in practice the techniques and capabilities of frame and object-oriented languages overlap significantly.
- Implementations:
  - KL-ONE,
  - LOOM [21],
  - PowerLoom [22],
  - OWL [23]
- Semantic editors
  - Protégé
- Semantic Reasoners (A semantic reasoner, rules engine, or simply a reasoner, is a piece of software able to infer logical consequences from a set of asserted facts or axioms.)
  - Pellet
  - RacerPro
  - FaCT++
  - HermiT



More representations

- Case Grammar
- Frames
- First Order Logic
- Semantic Nets
- Conceptual Dependency
- Rule-Based
- Conceptual Graphs



Ambiguity - Disambiguation

- I saw a man on a hill with a telescope meannings:
  - There’s a man on a hill, and I’m watching him with my telescope.
  - There’s a man on a hill, who I’m seeing, and *he* has a telescope.
  - There’s a man, and he’s on a hill that also has a telescope on it.
  - I’m on a hill, and I saw a man using a telescope.
  - There’s a man on a hill, and I’m sawing him with a telescope.

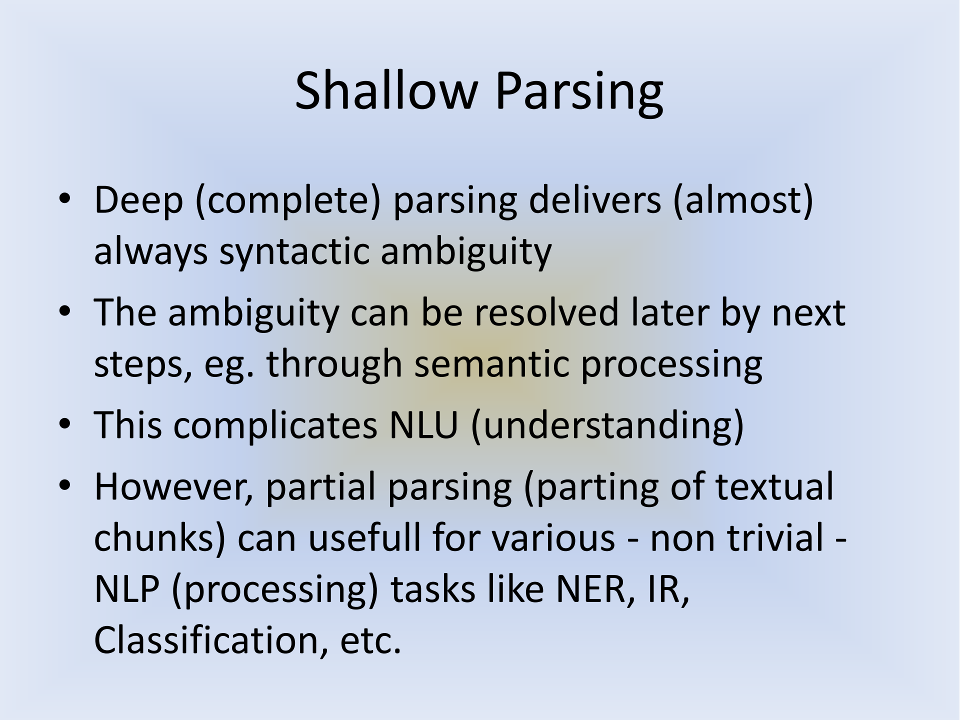
Syntactic ambiguity [6]

- Look at the dog with one eye. meanings:
    - – Look at the dog using only one of your eyes.
    - – Look at the dog that only has one eye.
- Both (this and previous) are syntactically ambiguous sentences

Syntactic Ambiguity – parsing Trees

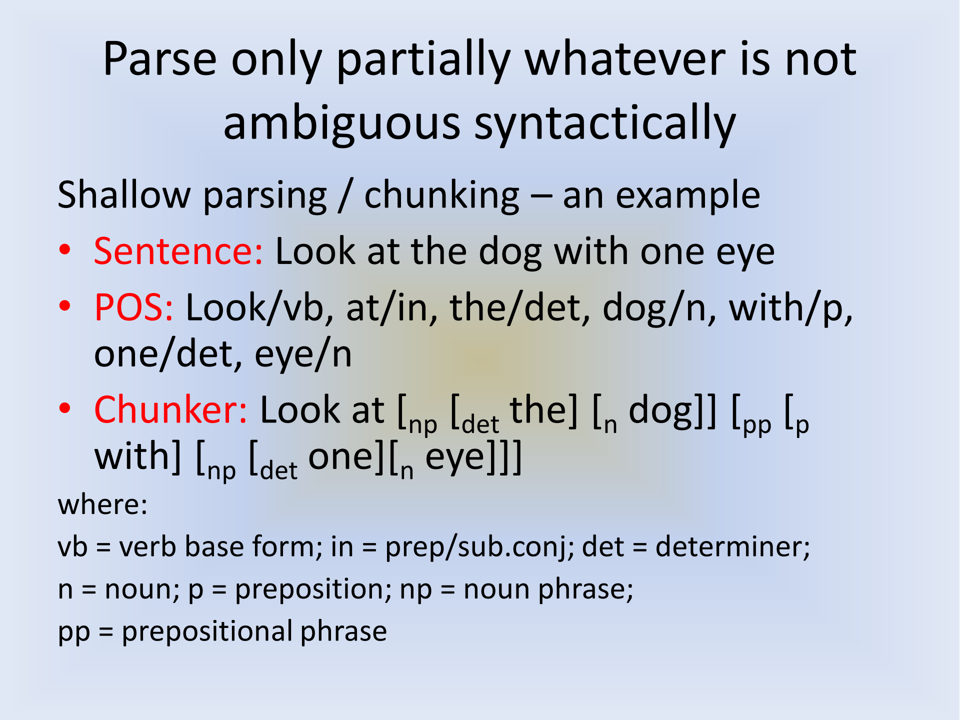
```
s( np(n(you))
   vp( v(look,present),
       pp  (p(at),np(det(the),n(dog))),
       pp  (p(with),np(...,n(eye)))
   )
)

s( np(n(you))
   vp( v(look,present),
       pp  (p(at), np( det(the),
                      np( n(dog),
                          pp( p(with),np(...,n(eye)))
                  )
       )
   )
)
```



Shallow Parsing

- Deep (complete) parsing delivers (almost) always syntactic ambiguity
- The ambiguity can be resolved later by next steps, eg. through semantic processing
- This complicates NLU (understanding)
- However, partial parsing (parting of textual chunks) can usefull for various - non trivial - NLP (processing) tasks like NER, IR, Classification, etc.



Parse only partially whatever is not ambiguous syntactically

Shallow parsing / chunking – an example

- Sentence: Look at the dog with one eye
- POS: Look/vb, at/in, the/det, dog/n, with/p, one/det, eye/n
- Chunker: Look at [$_{np}$ [$_{det}$ the] [$_{n}$ dog]] [$_{pp}$ [$_{p}$ with] [$_{np}$ [$_{det}$ one][$_{n}$ eye]]]

where:

vb = verb base form; in = prep/sub.conj; det = determiner;

n = noun; p = preposition; np = noun phrase;

pp = prepositional phrase

Three step process

- Word Identification (POS-Tagging)
- Chunk Identification (regular expressions OR Context Free Grammars OR Phrase Structure Grammars)
- Merging / Splitting of Chunks (via rules)
  - Combine adjacent chunks into a single chunk
  - *Define regular expressions that permit to merge sequences of adjacent chunks to a longer one*

Semantically Ambiguous

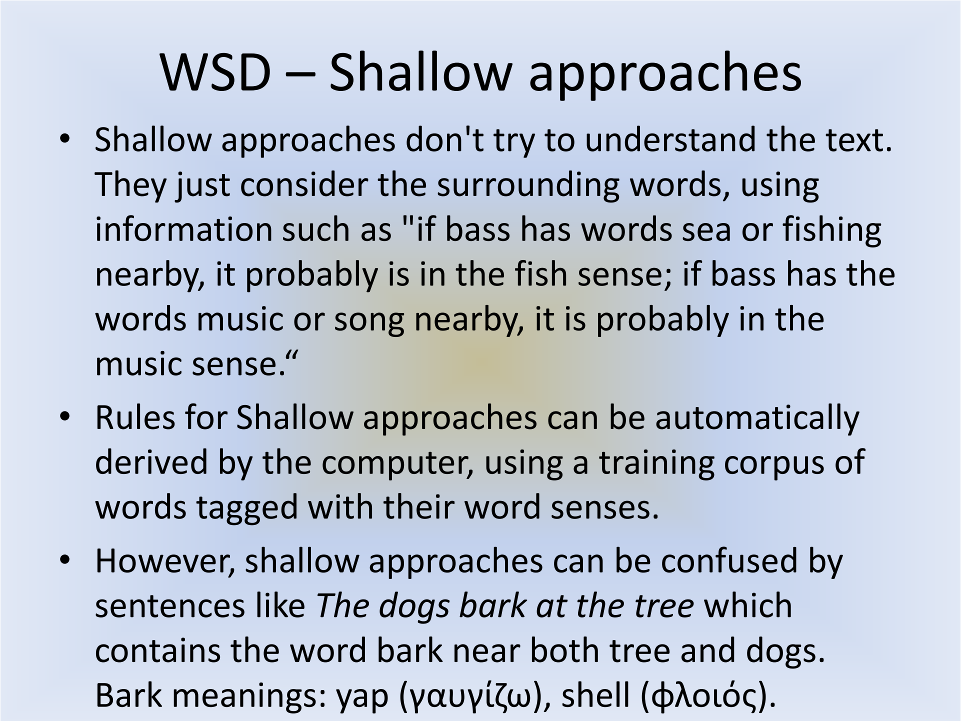
- word *slug* meanings:
  - Coin
  - Bullet
  - Loafer
  - Gastropod without shell
- word *bass* meanings:
  - a type of fish
  - tones of low frequency
  - a type of instrument
- Example sentences with ambiguity – word *bass*:
  - I went fishing for some sea *bass*
  - The *bass* line of the song is too weak

Word Sense Disambiguation (WSD) [7]

- word-sense disambiguation (WSD) is an open problem for natural language processing and ontology
- WSD is identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings
- The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference, *etc*.

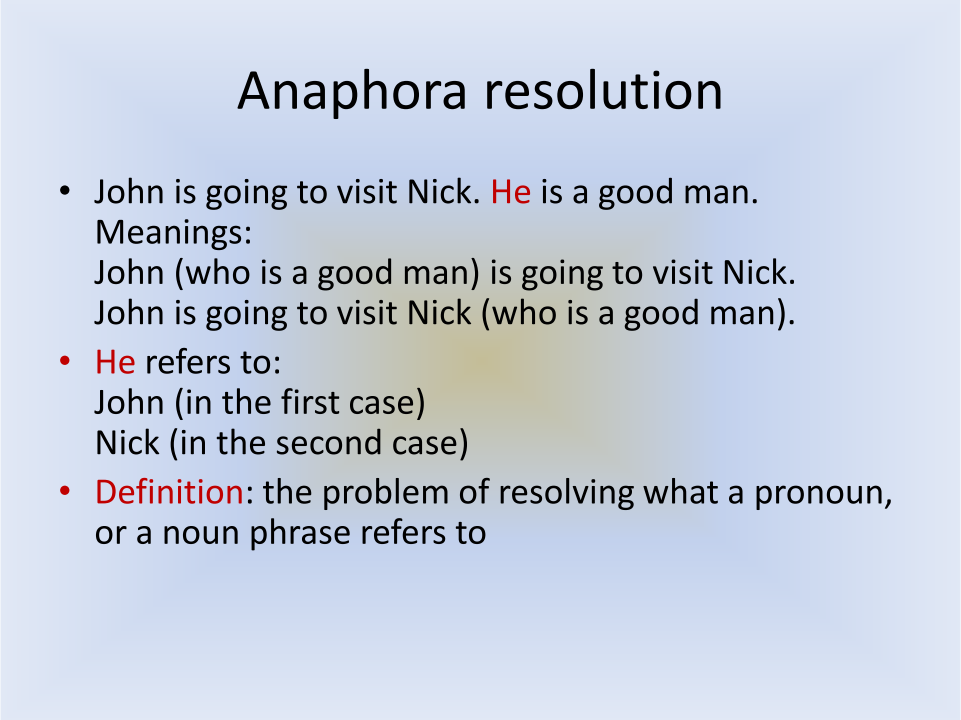
Word Sense Disambiguation (WSD) [7]

- As in all NLP there two main approaches for WSD
  - deep approaches
  - shallow approaches.
- Deep approaches presume access to a comprehensive body of world knowledge. Knowledge, such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish", is then used to determine in which sense the word bass is used. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside very limited domains.



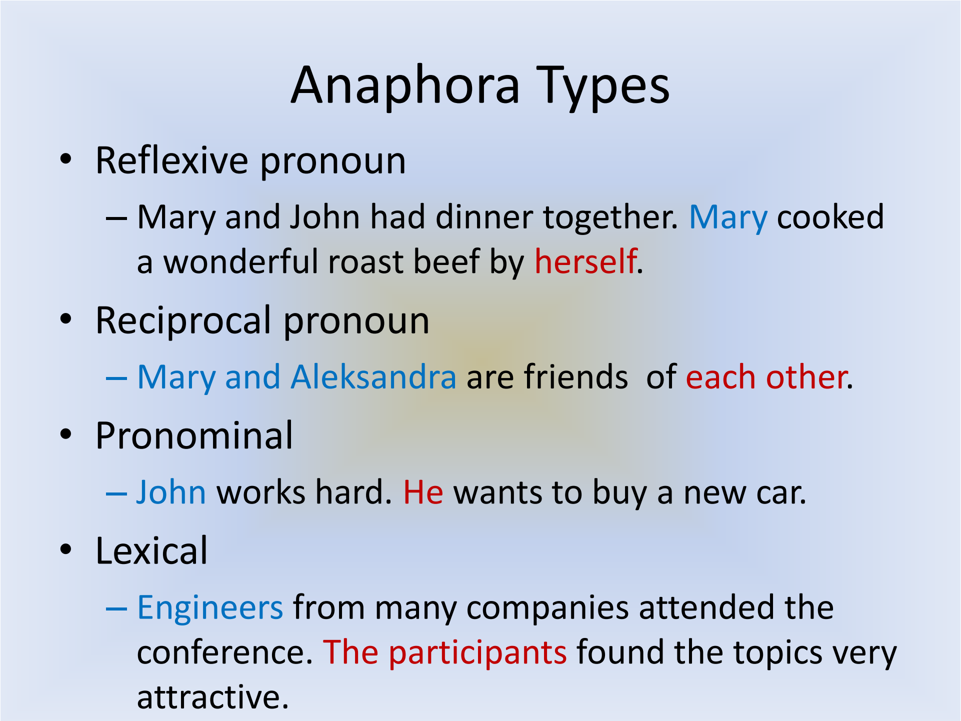
WSD – Shallow approaches

- Shallow approaches don't try to understand the text. They just consider the surrounding words, using information such as "if bass has words sea or fishing nearby, it probably is in the fish sense; if bass has the words music or song nearby, it is probably in the music sense.“
- Rules for Shallow approaches can be automatically derived by the computer, using a training corpus of words tagged with their word senses.
- However, shallow approaches can be confused by sentences like *The dogs bark at the tree* which contains the word bark near both tree and dogs. Bark meanings: yap (γαυγίζω), shell (φλοιός).



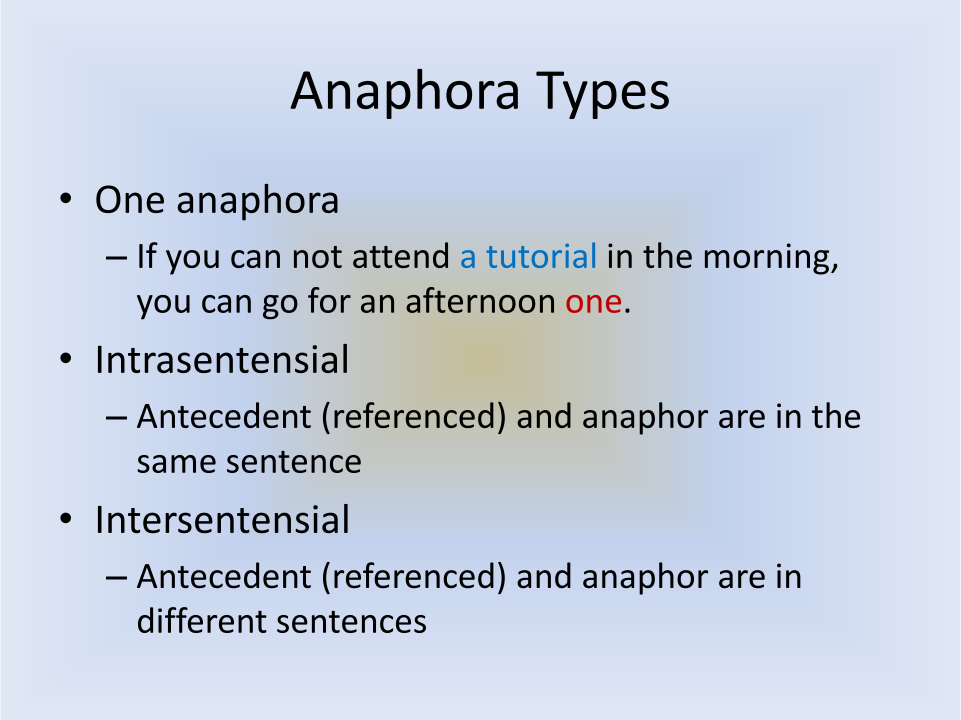
Anaphora resolution

- John is going to visit Nick. He is a good man.
  Meanings:
  John (who is a good man) is going to visit Nick.
  John is going to visit Nick (who is a good man).
- He refers to:
  John (in the first case)
  Nick (in the second case)
- Definition: the problem of resolving what a pronoun, or a noun phrase refers to



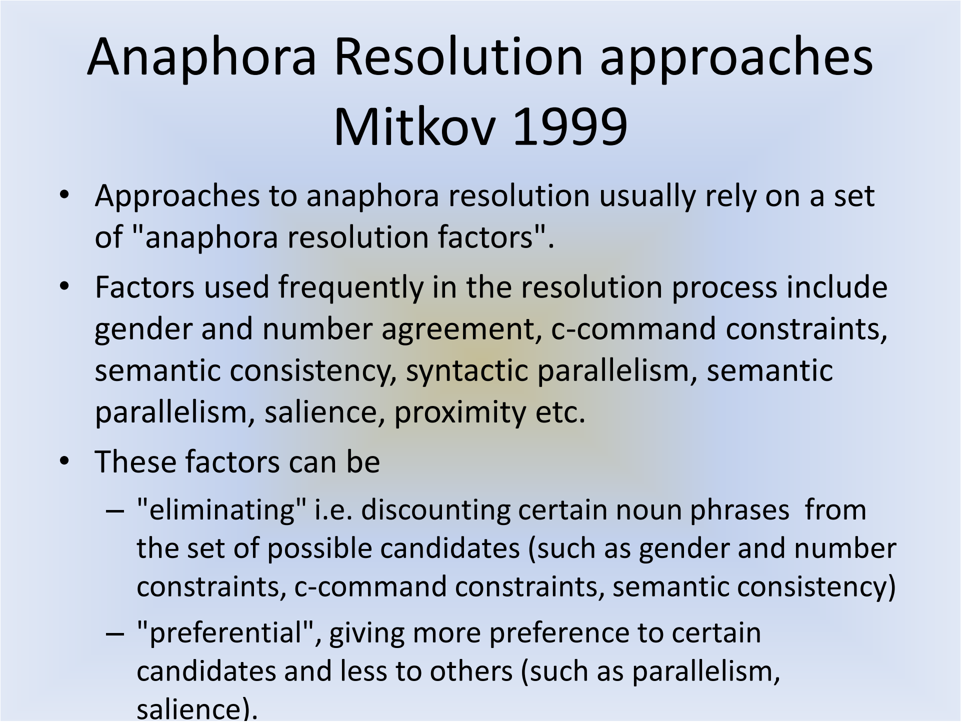
Anaphora Types

- Reflexive pronoun
  - Mary and John had dinner together. Mary cooked a wonderful roast beef by herself.
- Reciprocal pronoun
  - Mary and Aleksandra are friends of each other.
- Pronominal
  - John works hard. He wants to buy a new car.
- Lexical
  - Engineers from many companies attended the conference. The participants found the topics very attractive.



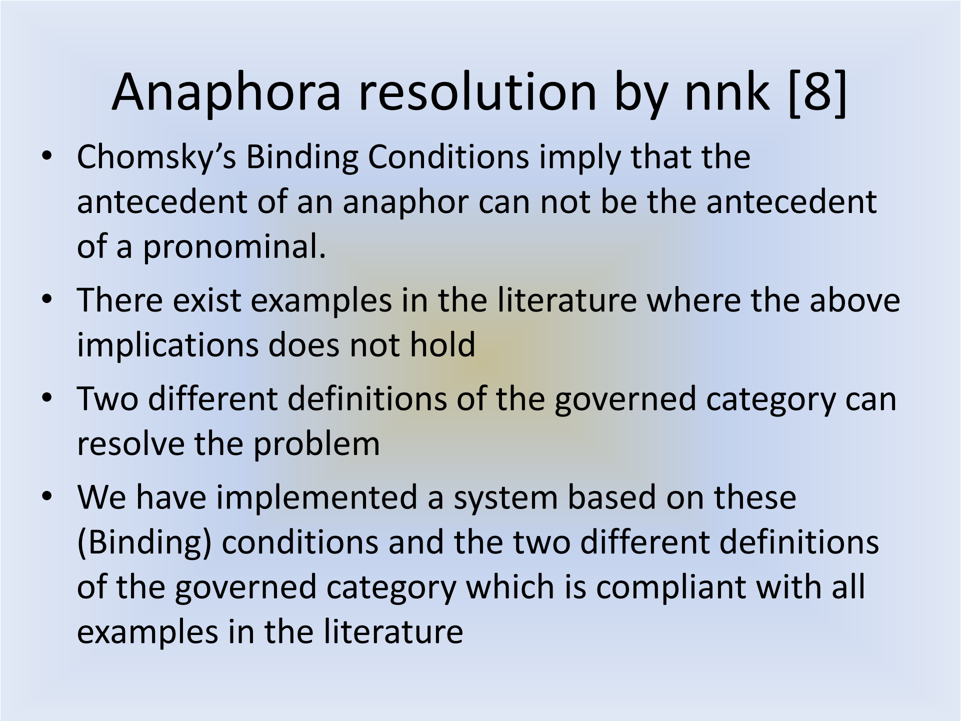
Anaphora Types

- One anaphora
  - If you can not attend a tutorial in the morning, you can go for an afternoon one.
- Intrasentensial
  - Antecedent (referenced) and anaphor are in the same sentence
- Intersentensial
  - Antecedent (referenced) and anaphor are in different sentences



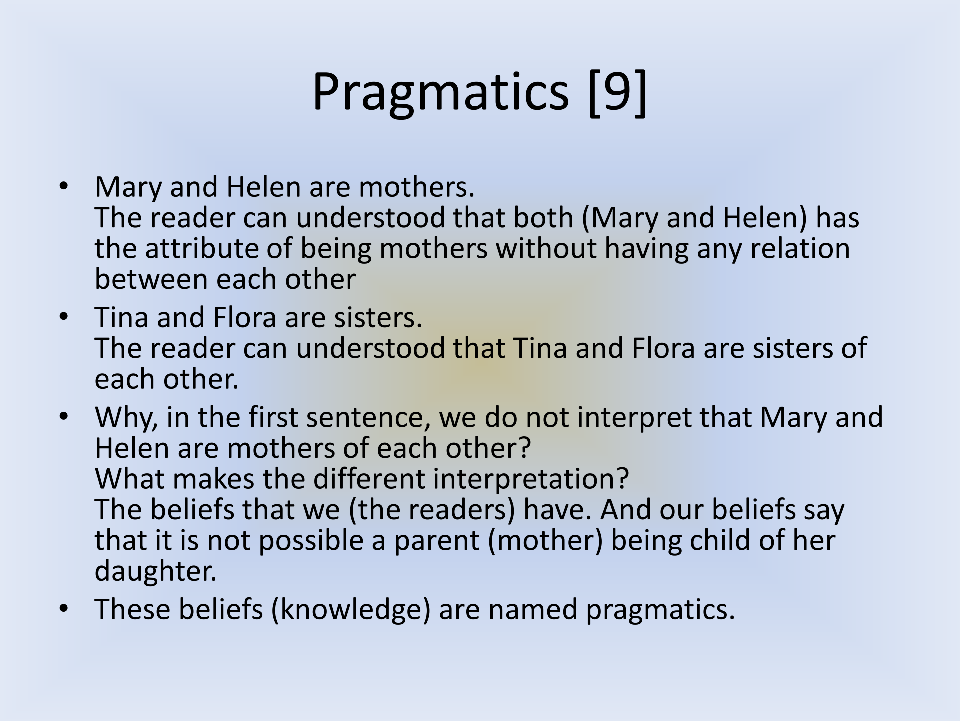
Anaphora Resolution approaches Mitkov 1999

- Approaches to anaphora resolution usually rely on a set of "anaphora resolution factors".
- Factors used frequently in the resolution process include gender and number agreement, c-command constraints, semantic consistency, syntactic parallelism, semantic parallelism, salience, proximity etc.
- These factors can be
  - "eliminating" i.e. discounting certain noun phrases from the set of possible candidates (such as gender and number constraints, c-command constraints, semantic consistency)
  - "preferential", giving more preference to certain candidates and less to others (such as parallelism, salience).



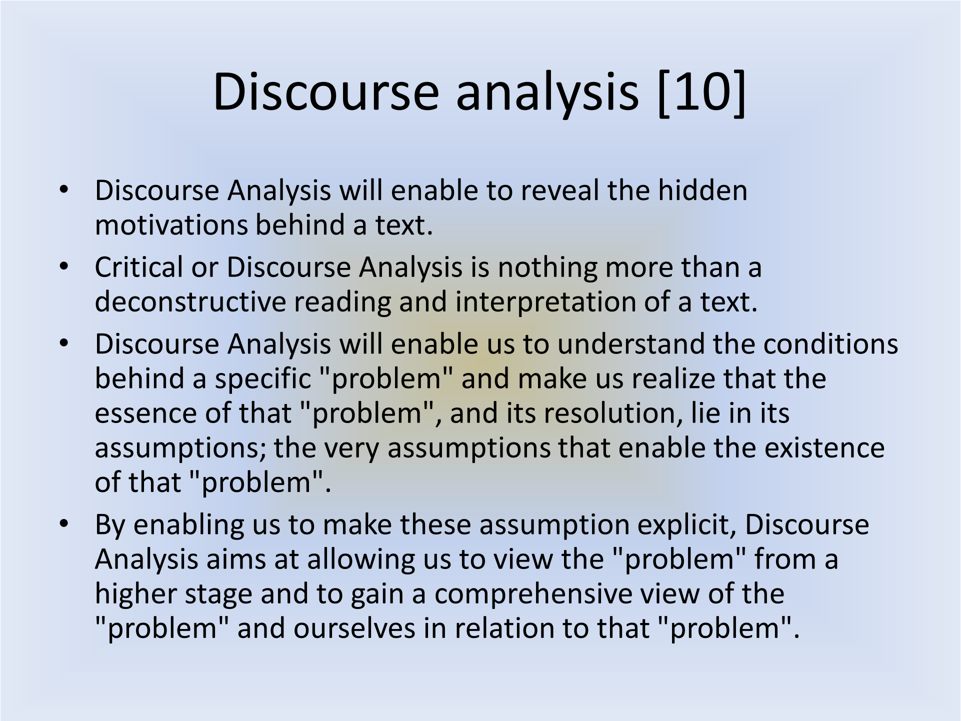
Anaphora resolution by nnk [8]

- Chomsky’s Binding Conditions imply that the antecedent of an anaphor can not be the antecedent of a pronominal.
- There exist examples in the literature where the above implications does not hold
- Two different definitions of the governed category can resolve the problem
- We have implemented a system based on these (Binding) conditions and the two different definitions of the governed category which is compliant with all examples in the literature



Pragmatics [9]

- Mary and Helen are mothers.
  The reader can understood that both (Mary and Helen) has the attribute of being mothers without having any relation between each other
- Tina and Flora are sisters.
  The reader can understood that Tina and Flora are sisters of each other.
- Why, in the first sentence, we do not interpret that Mary and Helen are mothers of each other?
  What makes the different interpretation?
  The beliefs that we (the readers) have. And our beliefs say that it is not possible a parent (mother) being child of her daughter.
- These beliefs (knowledge) are named pragmatics.



Discourse analysis [10]

- Discourse Analysis will enable to reveal the hidden motivations behind a text.
- Critical or Discourse Analysis is nothing more than a deconstructive reading and interpretation of a text.
- Discourse Analysis will enable us to understand the conditions behind a specific "problem" and make us realize that the essence of that "problem", and its resolution, lie in its assumptions; the very assumptions that enable the existence of that "problem".
- By enabling us to make these assumption explicit, Discourse Analysis aims at allowing us to view the "problem" from a higher stage and to gain a comprehensive view of the "problem" and ourselves in relation to that "problem".

References

- [1] Shef, What is Morphlogy
  http://all-about-linguistics.group.shef.ac.uk/branches-of-linguistics/morphology/what-is-morphology/

- [2] Wikipedia, Morphology
  https://en.wikipedia.org/wiki/Morphology_(linguistics)

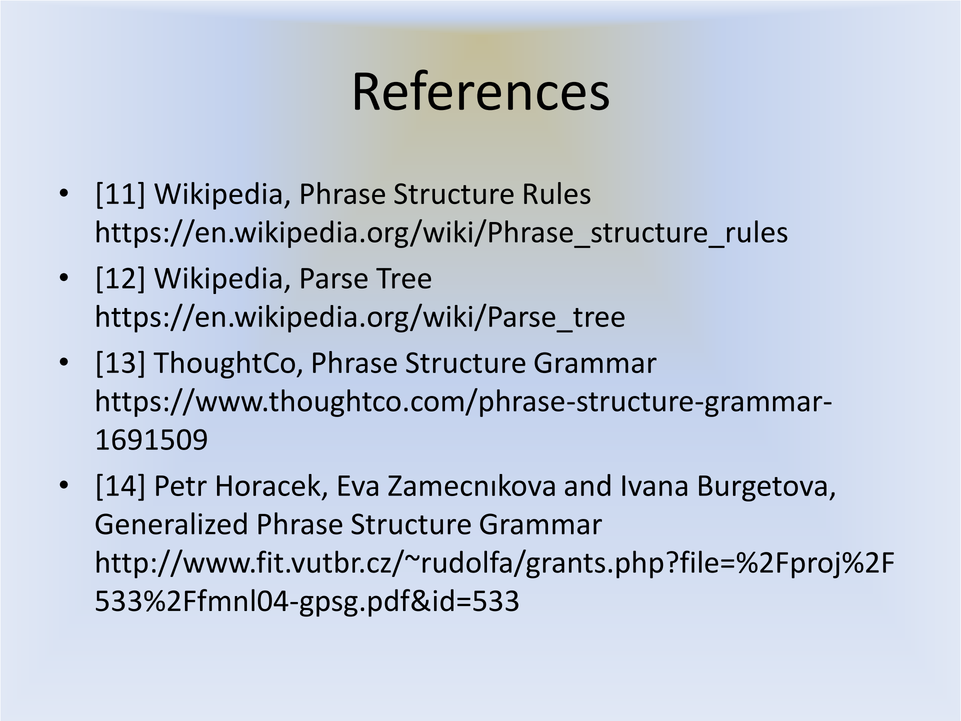
- [3] Wikipedia, Natural language understanding
  https://en.wikipedia.org/wiki/Natural_language_understanding

- [4] Wikipedia, Case Grammar
  https://en.wikipedia.org/wiki/Case_grammar

- [5] Fillmore grammatical cases - Perspective of Wallace Chafe
  http://linguistictheoryevolution.blogspot.al/2012/05/charles-fillmores-grammatical-cases.html

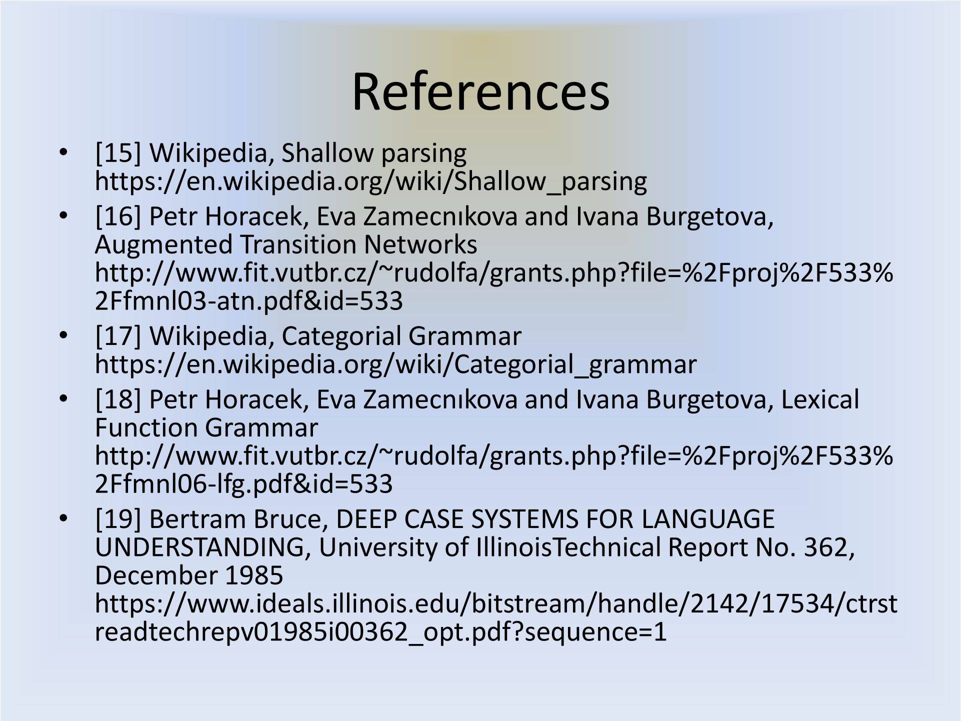
References

- [6] Linguistcs Online, Syntactic ambiguity
  http://languagelink.let.uu.nl/~lion/index.php?s=Grammar_exercises/grammar_4&lang=en
- [7] Wikipedia, Word-sense disambiguation
  https://en.wikipedia.org/wiki/Word-sense_disambiguation
- [8] Nikitas N. Karanikolas. Pronominal and Anaphor Resolution. Computing and Information Technology (CIT) journal, volume 1, No 3, 1993.
  http://users.teiath.gr/nnk/papers/A01_scan.pdf
- [9] Wikipedia, Pragmatics
  https://en.wikipedia.org/wiki/Pragmatics
- [10] Discourse Analysis: A resource book for students.
  http://routledgetextbooks.com/textbooks/_author/9780415610001-jones/



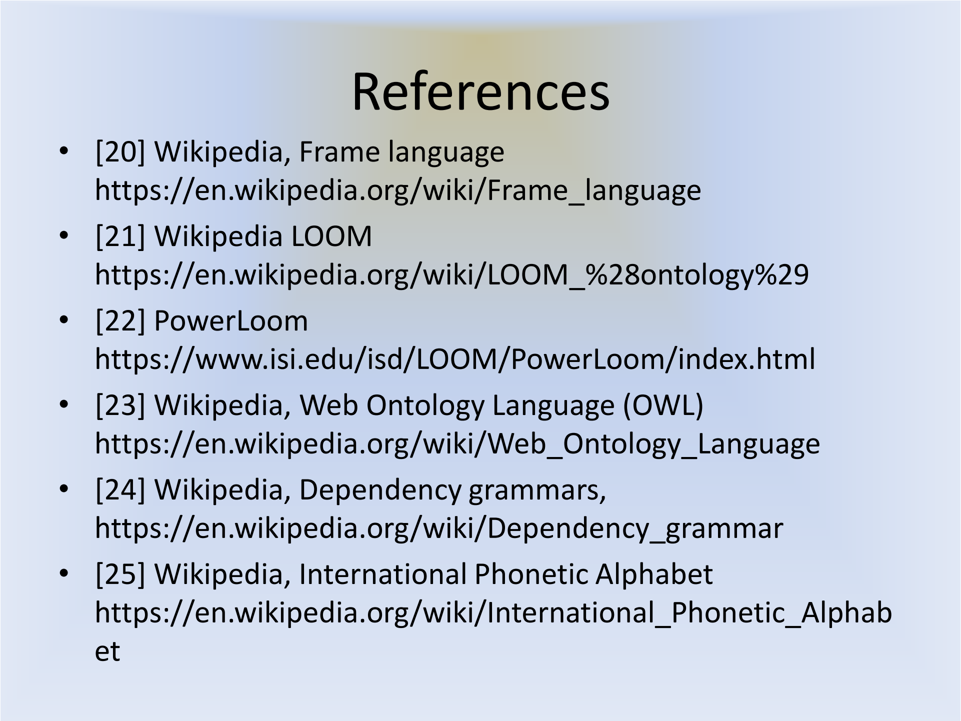
References

- [11] Wikipedia, Phrase Structure Rules https://en.wikipedia.org/wiki/Phrase_structure_rules
- [12] Wikipedia, Parse Tree https://en.wikipedia.org/wiki/Parse_tree
- [13] ThoughtCo, Phrase Structure Grammar https://www.thoughtco.com/phrase-structure-grammar-1691509
- [14] Petr Horacek, Eva Zamecnıkova and Ivana Burgetova, Generalized Phrase Structure Grammar http://www.fit.vutbr.cz/~rudolfa/grants.php?file=%2Fproj%2F533%2Ffmnl04-gpsg.pdf&id=533



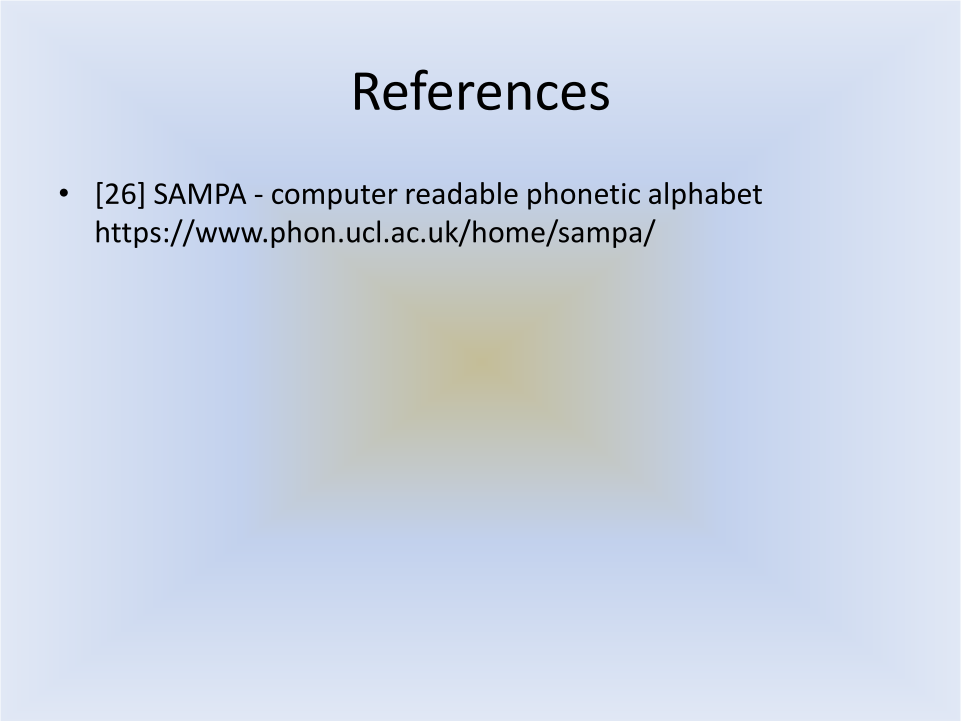
References

- [15] Wikipedia, Shallow parsing https://en.wikipedia.org/wiki/Shallow_parsing
- [16] Petr Horacek, Eva Zamecnıkova and Ivana Burgetova, Augmented Transition Networks http://www.fit.vutbr.cz/~rudolfa/grants.php?file=%2Fproj%2F533%2Ffmnl03-atn.pdf&id=533
- [17] Wikipedia, Categorial Grammar https://en.wikipedia.org/wiki/Categorial_grammar
- [18] Petr Horacek, Eva Zamecnıkova and Ivana Burgetova, Lexical Function Grammar http://www.fit.vutbr.cz/~rudolfa/grants.php?file=%2Fproj%2F533%2Ffmnl06-lfg.pdf&id=533
- [19] Bertram Bruce, DEEP CASE SYSTEMS FOR LANGUAGE UNDERSTANDING, University of IllinoisTechnical Report No. 362, December 1985 https://www.ideals.illinois.edu/bitstream/handle/2142/17534/ctrstreadtechrepv01985i00362_opt.pdf?sequence=1



References

- [20] Wikipedia, Frame language https://en.wikipedia.org/wiki/Frame_language
- [21] Wikipedia LOOM https://en.wikipedia.org/wiki/LOOM_%28ontology%29
- [22] PowerLoom https://www.isi.edu/isd/LOOM/PowerLoom/index.html
- [23] Wikipedia, Web Ontology Language (OWL) https://en.wikipedia.org/wiki/Web_Ontology_Language
- [24] Wikipedia, Dependency grammars, https://en.wikipedia.org/wiki/Dependency_grammar
- [25] Wikipedia, International Phonetic Alphabet https://en.wikipedia.org/wiki/International_Phonetic_Alphabet



References

- [26] SAMPA - computer readable phonetic alphabet https://www.phon.ucl.ac.uk/home/sampa/