10 CÁTEDRA INTERNACIONAL DE INGENIERÍA
INGENIERÍA Y PAZ

Introduction to NLP

Natural Language Processing and Text Mining, summer school 2016

Ing. Sergio Jiménez, Ms. C., Ph. D.

Outline

- What is Natural Language Processing?
- General approaches
- Tasks addressed by NLP

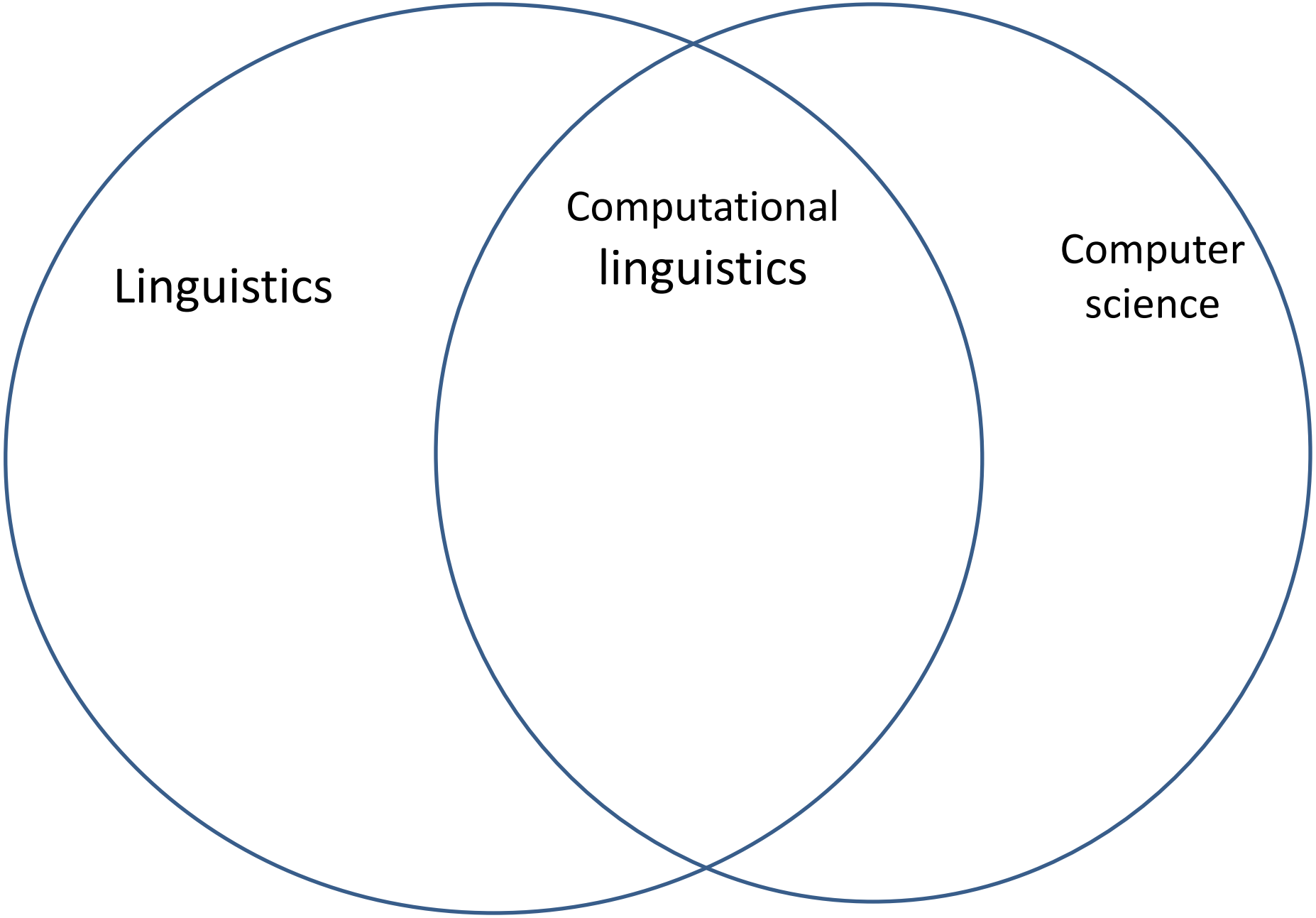
Outline

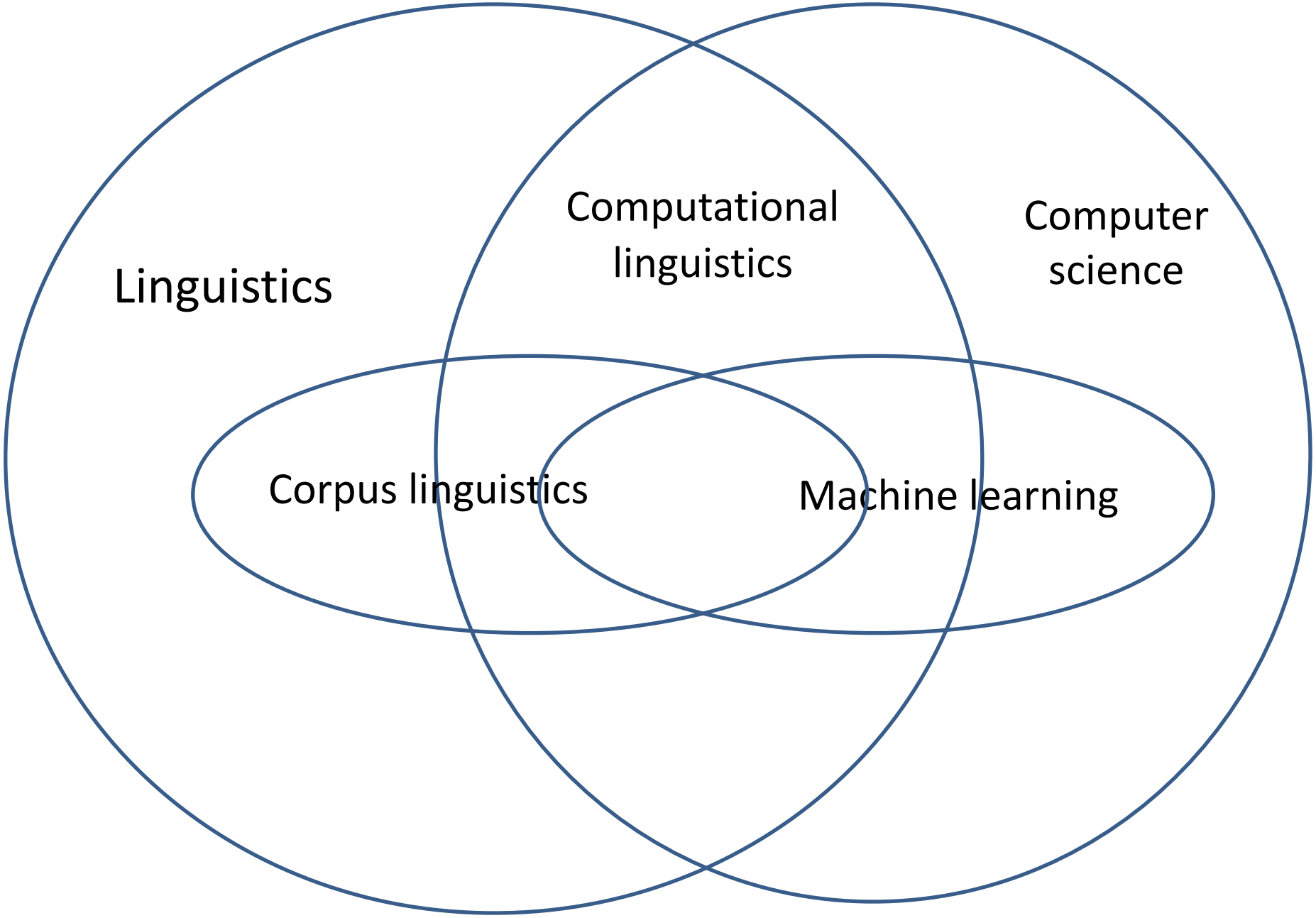
- What is Natural Language Processing?
- General approaches
- Tasks addressed by NLP

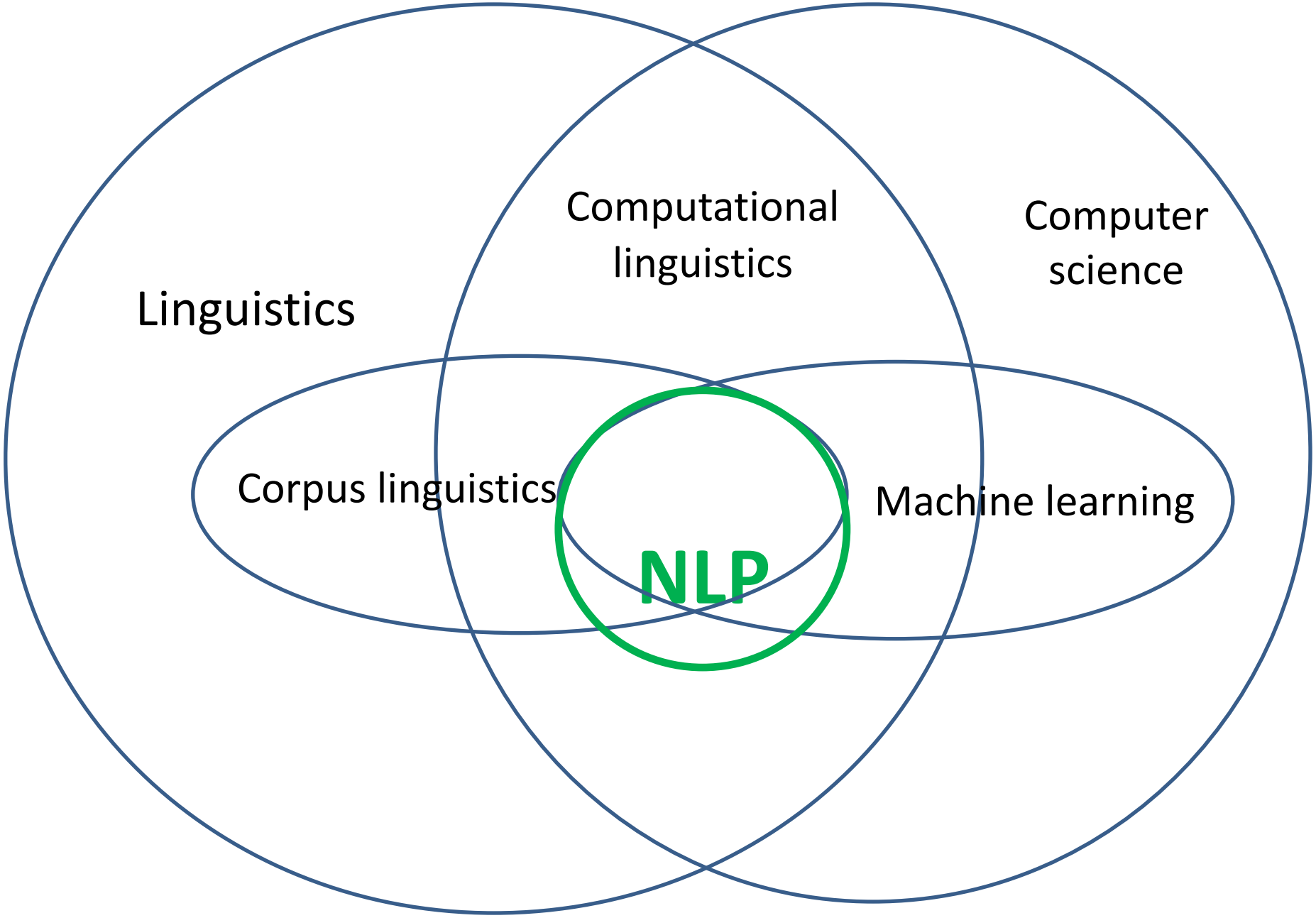
What is NLP?

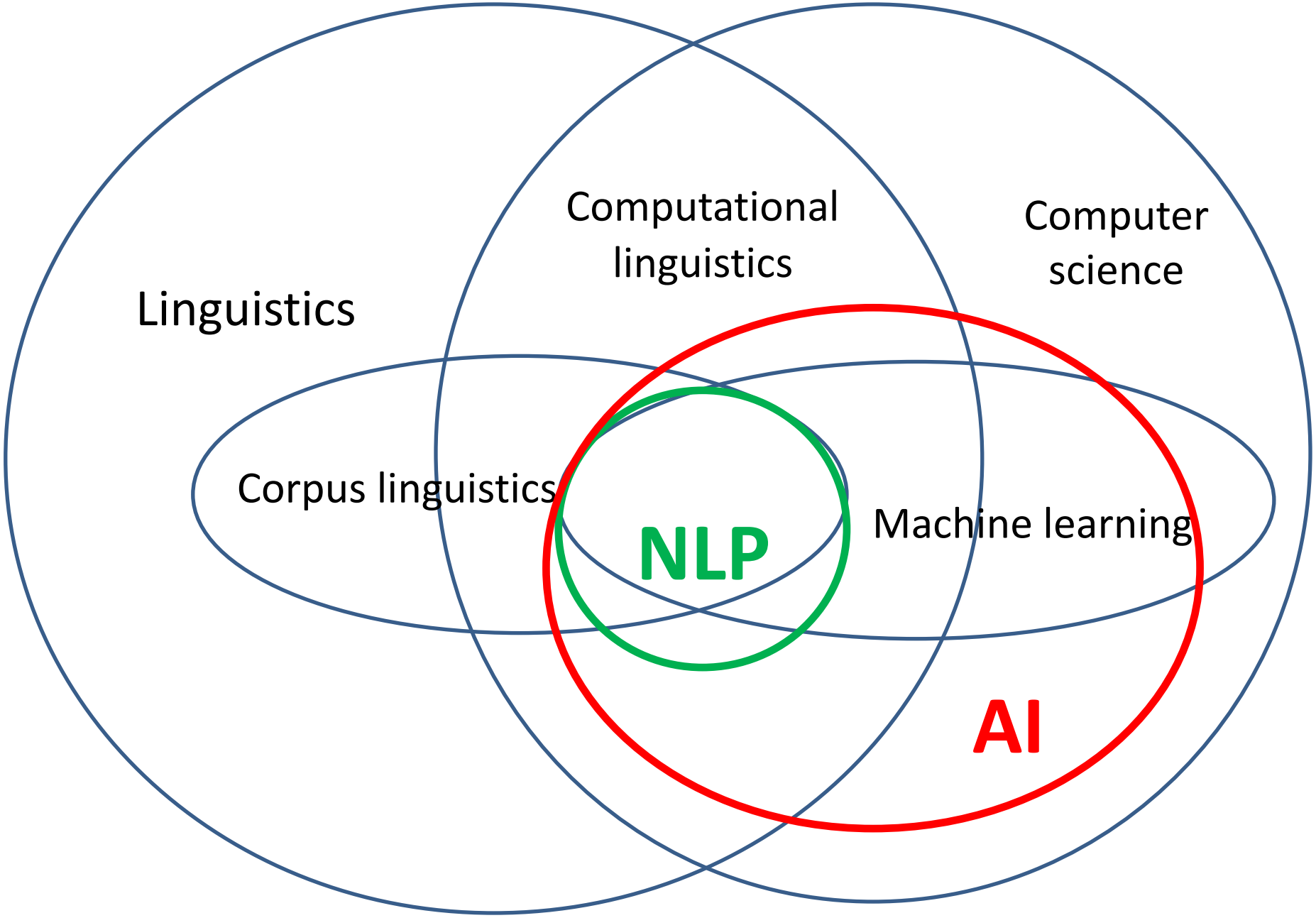
- Natural language processing

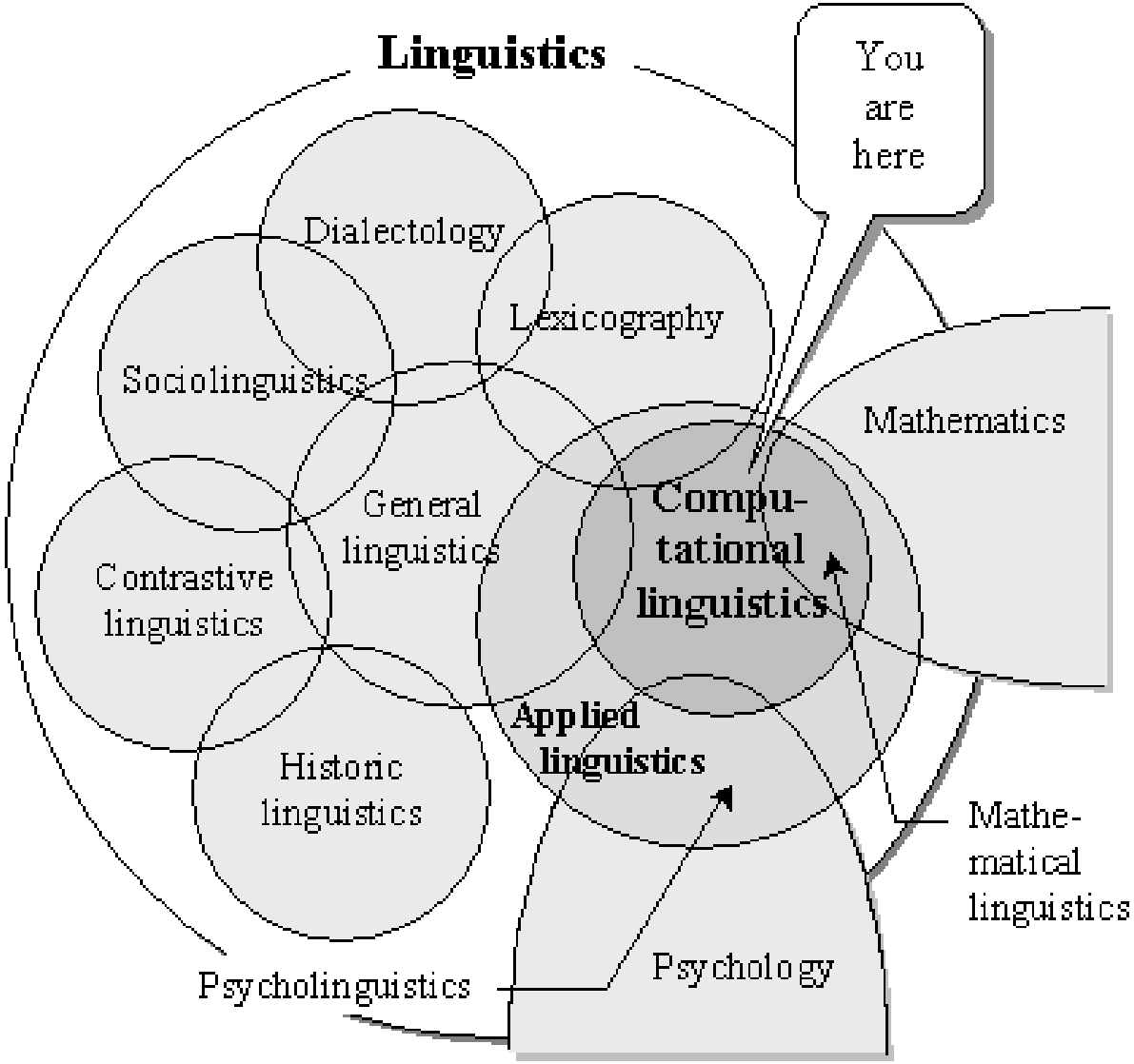
- Wikipedia's definition: "(NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.


Linguistics
Computational linguistics
Computer science


Linguistics
Computational linguistics
Computer science
Corpus linguistics
Machine learning


Computational linguistics
Computer science
Linguistics
Corpus linguistics
Machine learning
NLP


Computational linguistics
Computer science
Linguistics
Corpus linguistics
NLP
Machine learning
AI



Taken from “COMPUTATIONAL LINGUISTICS: Models, Resources and Applications” by Bolshakov and Gelbukh, 2004

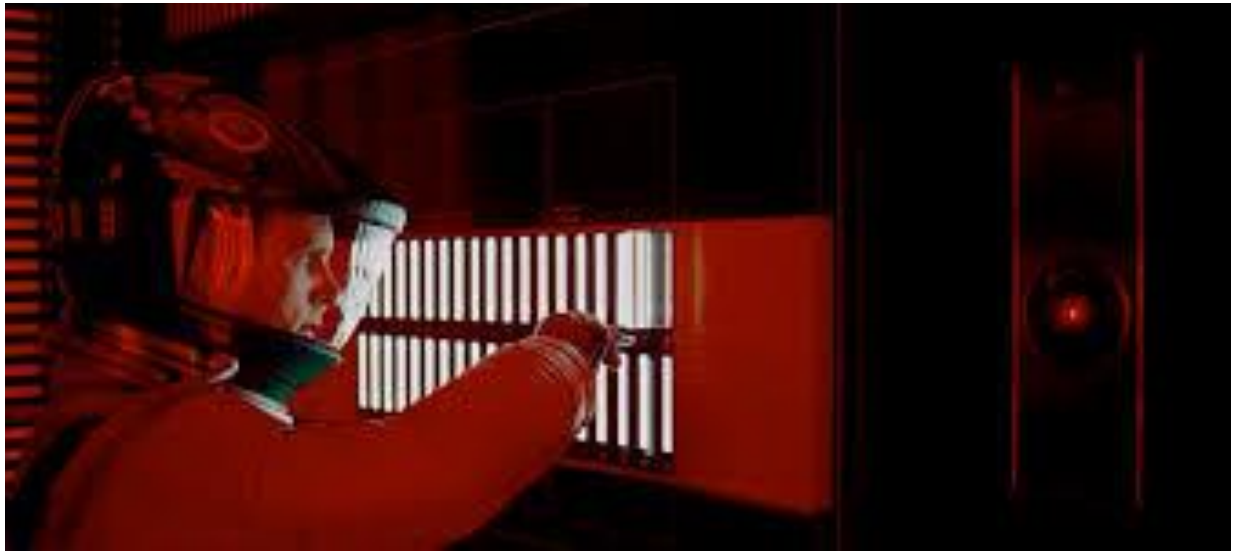
What is NLP's ultimate goal?

- "To make machines understands human language", A. Gelbukh

- "The goal of the Natural Language Processing (NLP) ~~group~~ is to design and build software that will analyze, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person." , http://research.microsoft.com/en-us/groups/nlp/

NLP in popular Science Fiction

“2001: A Space Odyssey” (1968, S. Kubrick)



HAL 9000 (Heuristically programmed Algoritmic computer)

NLP in popular Science Fiction

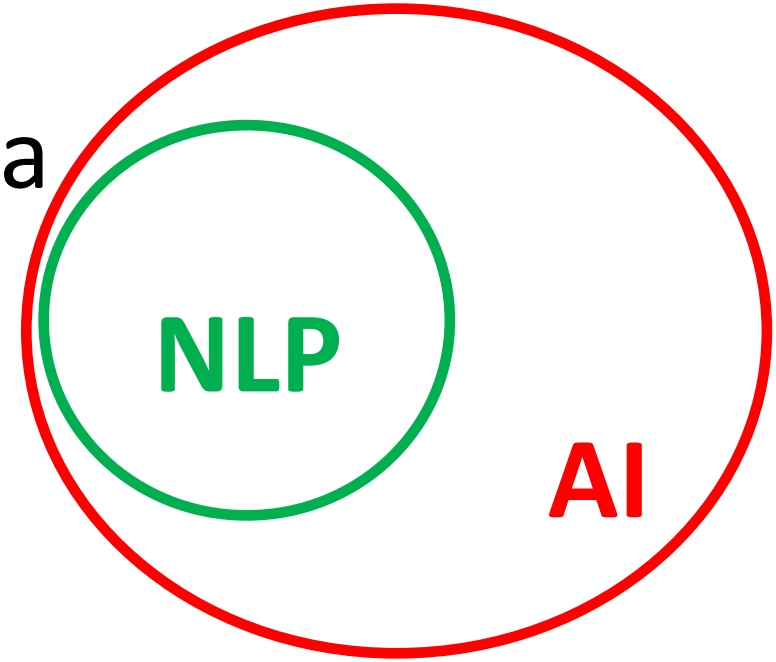
"Robot and Frank" (2012, J. Schreier)

NLP's extrinsic goal: make computing accessible to everybody.

Outline

- What is Natural Language Processing?
- General approaches
- Tasks addressed by NLP

AI approach and dilemma



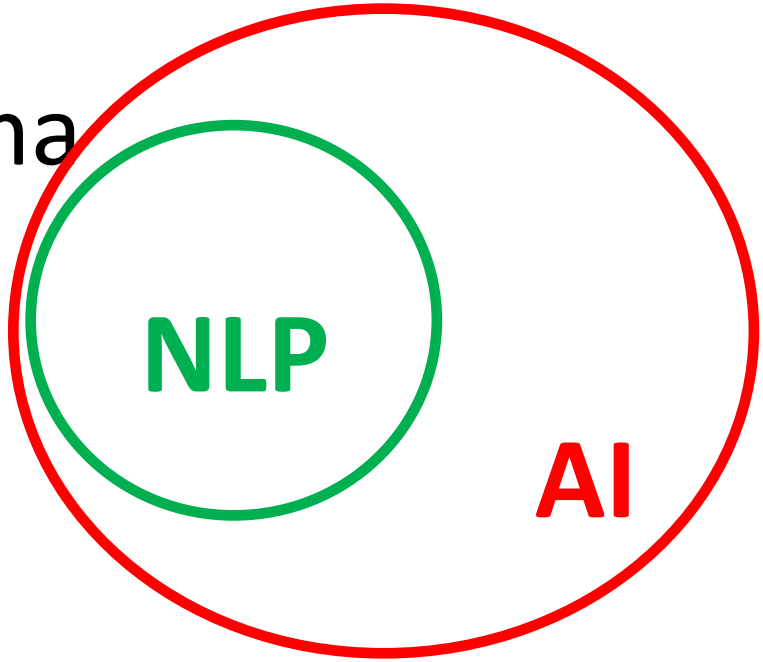
Understand how we think to …

… build intelligent entities.

Or

See how intelligent entities behave to …

NLP approach and dilemma



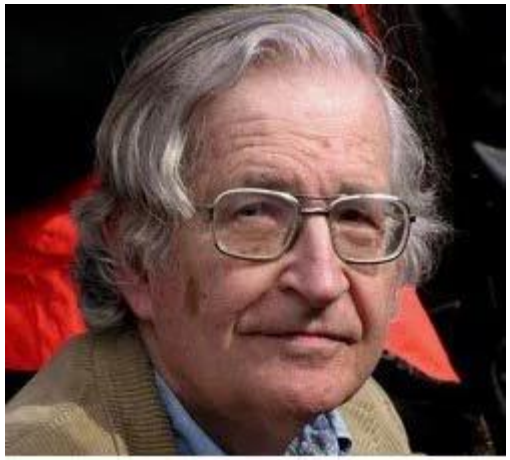
Understand how humans understand and produce language to …

… build systems that understand and generate natural language.

Or

See how language is used by humans to …

Norvig vs. Chomsky and the Fight for the Future of AI



Rule-based models

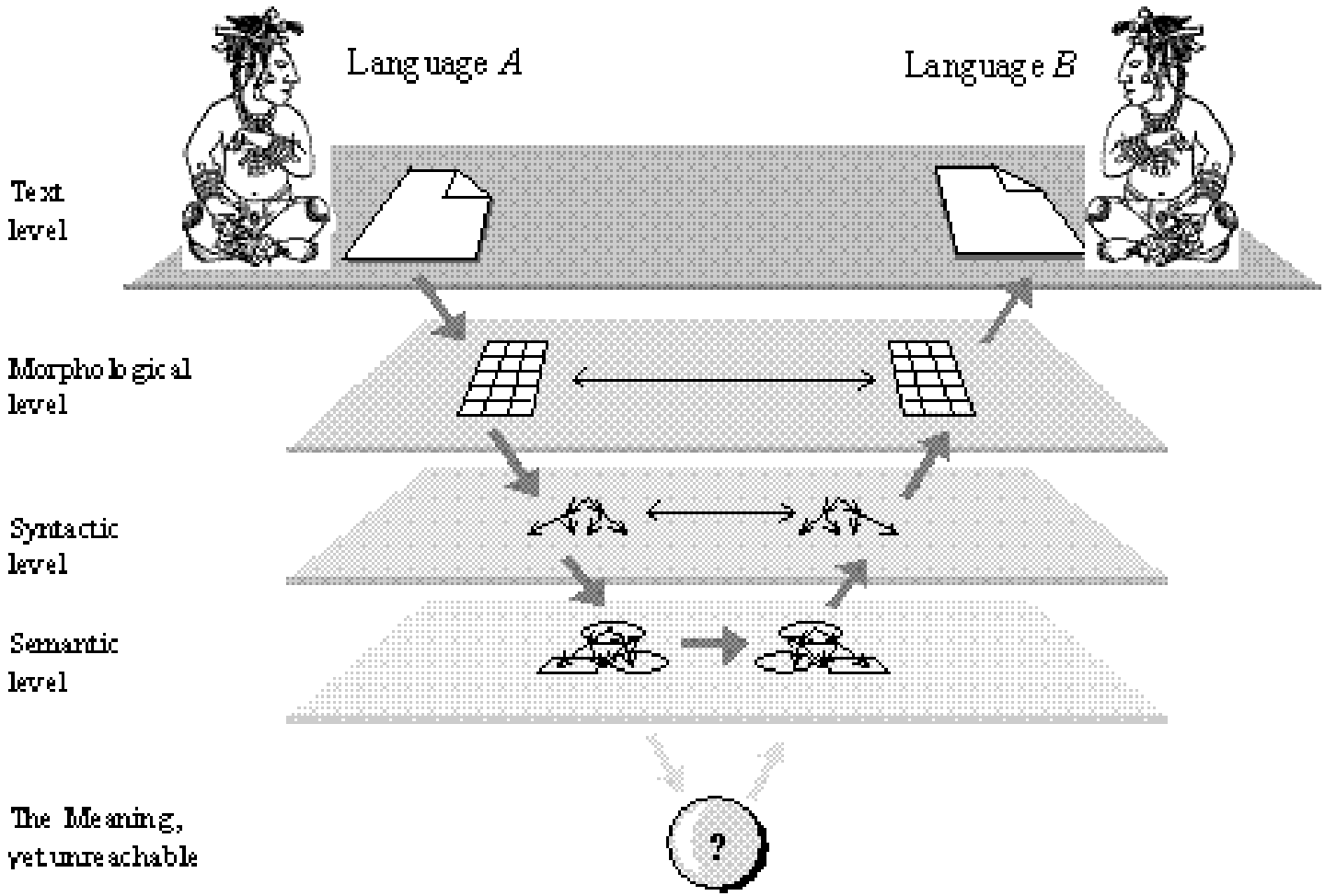
"… purely statistical methods to produce behavior that mimics something in the world."

Data-driven models

"… with enough data, attempting to fit any simple model at all is pointless."

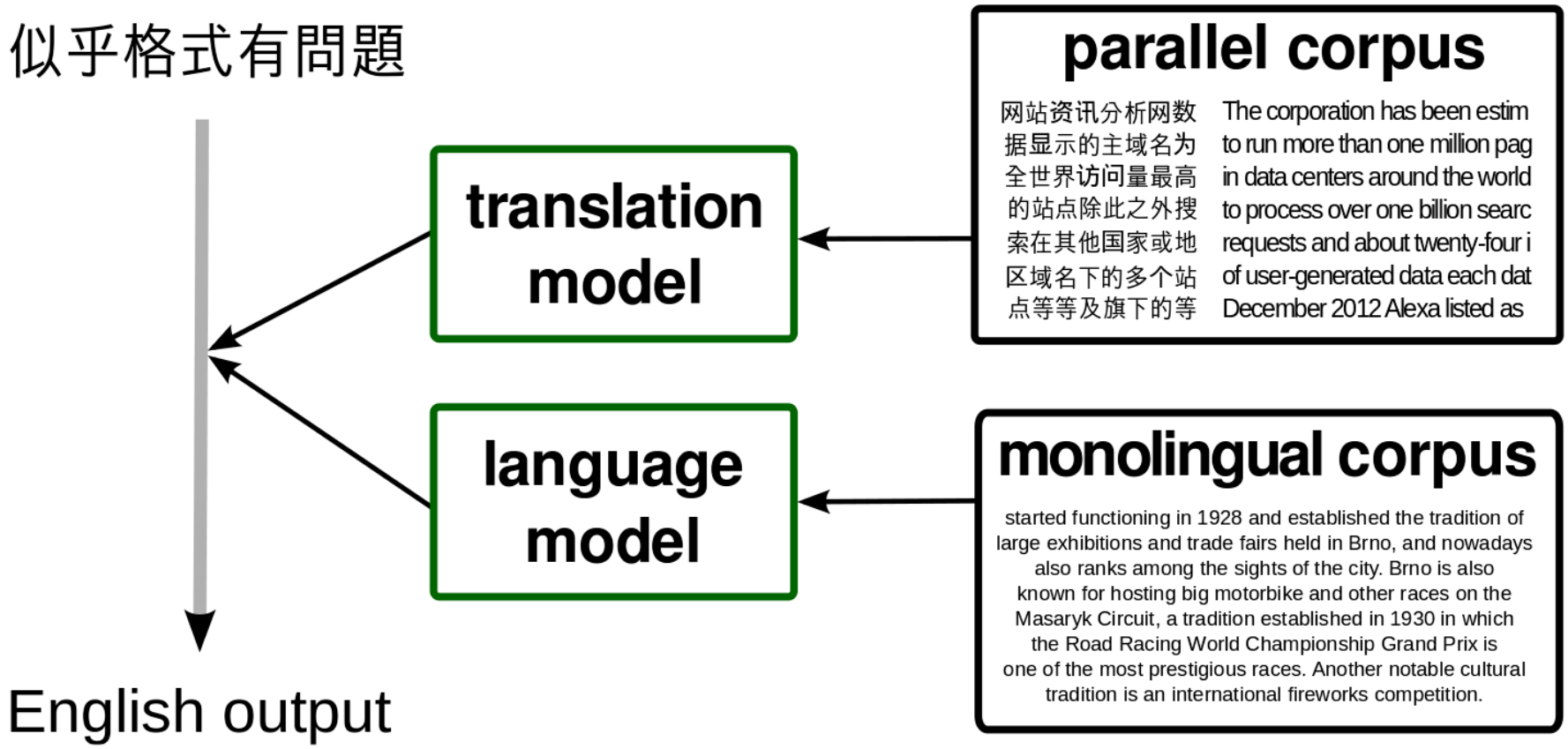
http://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/

Translation by understanding



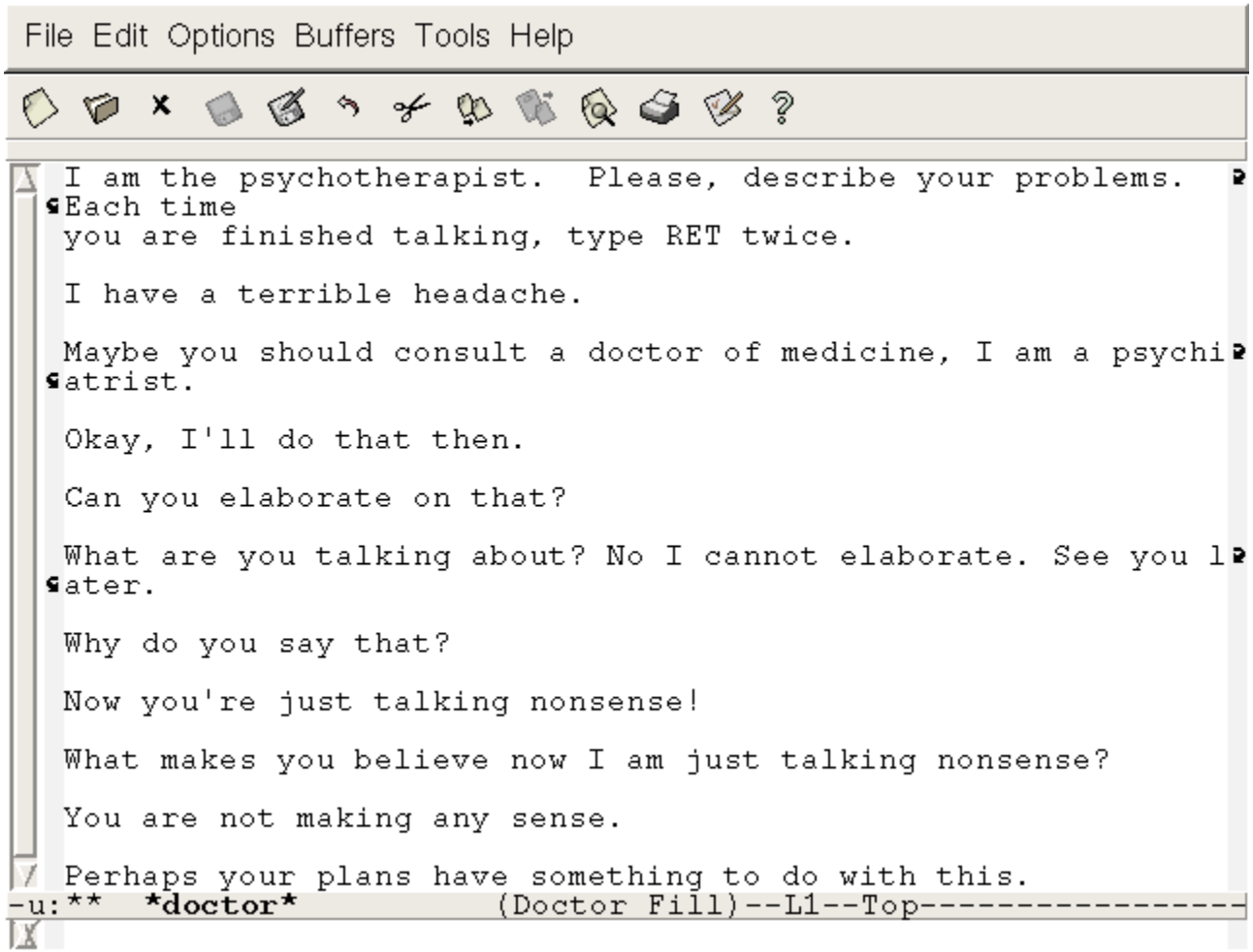
Taken from “COMPUTATIONAL LINGUISTICS: Models, Resources and Applications” by Bolshakov and Gelbukh, 2004

Translation by statistics



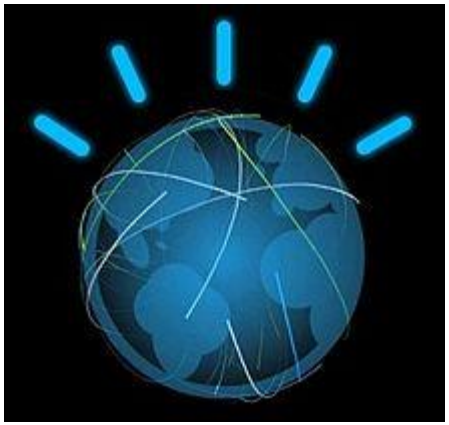
https://nlp.fi.muni.cz/web3/en/MachineTranslation

Some history
1965 ELIZA



http://www.manifestation.com/neurotoys/eliza.php3

IBM’s Watson at Jeopardy! (2011)

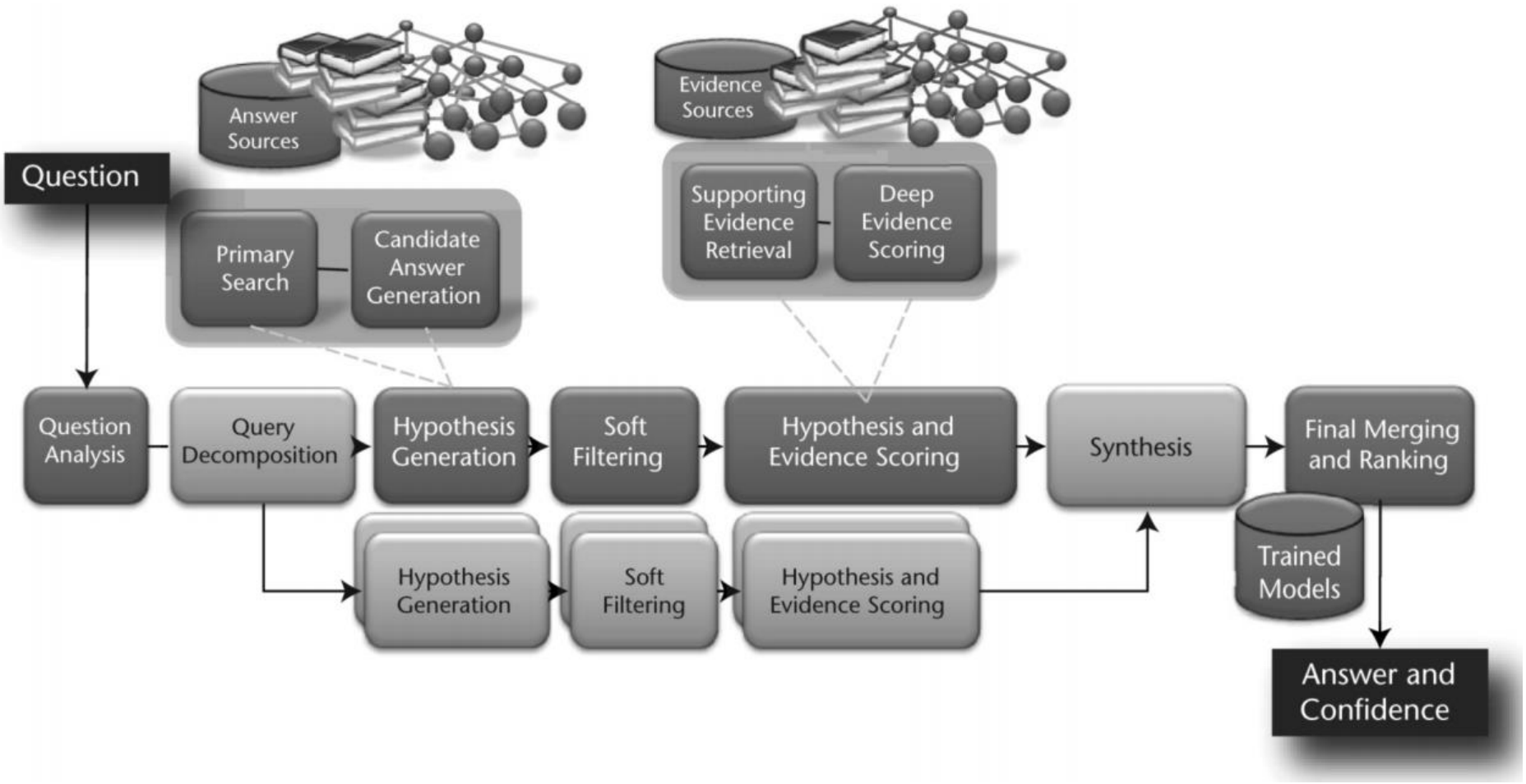


Statistics

Development Team	25 people
Project Duration	4 years
Software	1,000,000+ SLOC
	700K Java, 300K C++, plus other bits
	~ 130 components
Hardware	90 IBM Power 750 servers
	2880 Power7 cores @ 80+ TFLOPS
	20 TB memory
	10 Gbps network

WITH MUCH "GRAVITY", THIS YOUNG FELLOW OF TRINITY BECAME THE LUCASIAN PROFESSOR OF MATHEMATICS IN 1669

Watson's architecture



Taken from "Building Watson: An Overview of the DeepQA Project" by Ferrucci et al., AI Magazine, Fall 2010. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303/2165

How much knowledge from linguistics is needed for NLP R&D?

Igor Boshakov

“ How dare to investigate in NLP with only knowledge of high school linguistics! ” CICLING 2009

Linguistics resources needed for semantic textual similarity

Year	Methods used by the three top systems	Resources used
2012	Feature-based ML, Similarity functions, Soft Cardinality, ESA, SMT	String matching, KB similarity, Wikipedia, Wikitionary, BIUTEE textual entailment, distributional thesaurus, multilingual corpora
2013	Feature-based ML, feature selection, LSA, word alignment, LDA	2012 resources +: WordNet, WebBase 2007, POS tagger, Wikitionary
2014	Word alignments, feature-based ML, Soft cardinality, LSA/LSI	String similarity, NER, dependency parsing, PPDB, Wordnik
2015	Word alignments, feature-based ML	Word embedding, PPDB, POS tagging, WordNet
2016	Convolutional neural network (UMD 5th /40 team)	Sentence embeddings

Outline

- What is Natural Language Processing?
- General approaches
- Tasks addressed by NLP

Task addressed by NLP (main categories)

1. Reveal hidden structure of the language, e.g. POS tagging, parsing, stemming
2. Perform “atomic” tasks of “understanding” in text, e.g. WSD, paraphrase detection, TE
3. Perform simple but tedious task involving large amount of texts, e.g. IR, QA, IE, summarization, etc.

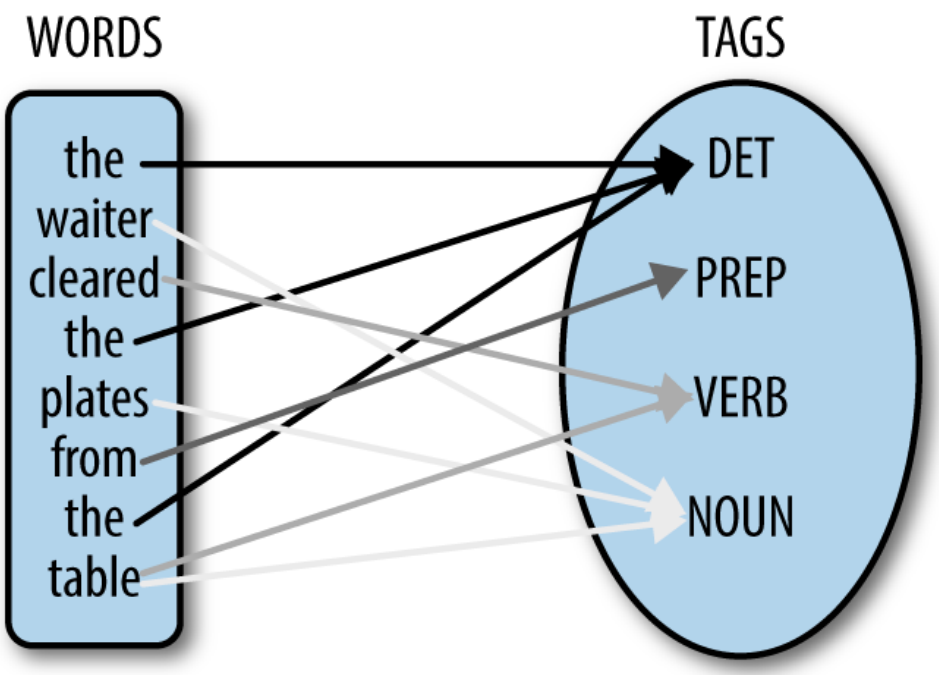
Task addressed by NLP (main categories)

1. Reveal hidden structure of the language, e.g. POS tagging, parsing, stemming
2. Perform “atomic” tasks of “understanding” in text, e.g. WSD, paraphrase detection, TE
3. Perform simple but tedious task involving large amount of texts, e.g. IR, QA, IE, summarization, etc.

1. Reveal hidden structure of the language
POS tagging

"white"

Noun: piece of laundry, white part of an egg, etc.
Adjective: color white
Verb: to cover with white coloring



Some POS tags used in English

AT	article
BEZ	the word *is*
IN	preposition
JJ	adjective
JJR	comparative adjective
MD	modal (*may, can, ...*)
MN	singular or mass noun
NNP	singular proper noun
NNS	plural noun
PERIOD	. : ? !
PN	personal pronoun
RB	adverb
RBR	comparative adverb
TO	the word *to*
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle
VBN	verb, past participle
VBP	verb, non 3d person singular present
VBZ	verb, 3d person singular present
WDT	wh-determiner (*what, which ...*)

http://slideplayer.com/slide/6426089/

Performance of POS taggers

- Most successful algorithms disambiguate about 96%-97% of the tokens!
- Information of taggers is quite usefull for information extraction, question answering and shallow parsing.

http://slideplayer.com/slide/6426089/

POS taggers

- Methods: Hidden Markov Models, Conditional Random Fields, Machine Learning classifiers

- Popular taggers:
  - Stanford POS tagger http://nlp.stanford.edu/software/tagger.html
  - TreeTagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
  - NLTK POS tagger http://www.nltk.org/book/ch05.html

1. Reveal hidden structure of the language
Shallow Parsing

He	*reckons*	*the current account deficit*	*will narrow*	*to*	
NP	VP	NP	VP	PP	
only # 1.8 billion	*in*	*September* .			
NP	PP	NP			

A chunker (shallow parser) segments a sentence into meaningful phrases.

Source: personalpages.manchester.ac.uk/staff/Sophia.Ananiadou/**DTCII**.ppt

Chunkers and Sallow Parsing tools

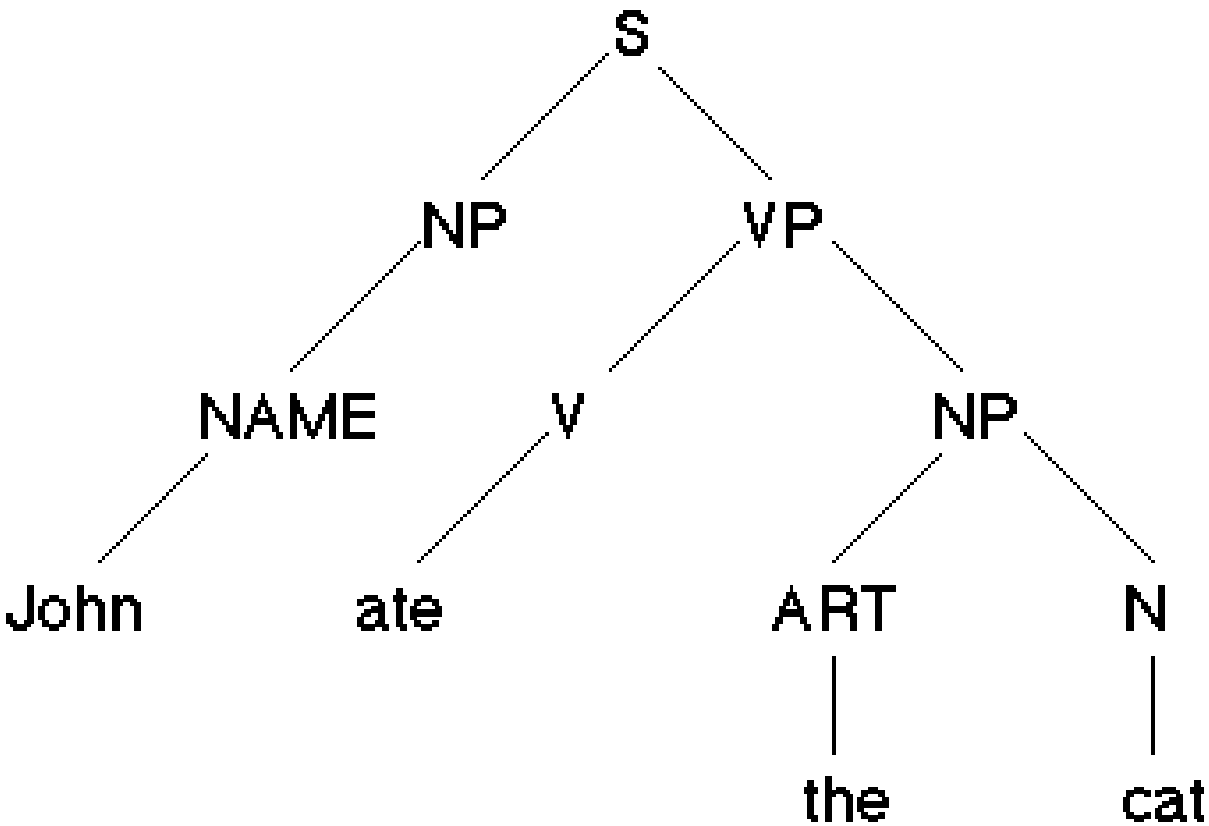
Popular chunkers and shallow parsers:

- Apache OpenNLP OpenNLP includes a chunker.
- GATE General Architecture for Text Engineering GATE includes a chunker.
- NLTK chunking
- Illinois Shallow Parser Shallow Parser Demo

https://en.wikipedia.org/wiki/Shallow_parsing

1. Reveal hidden structure of the language
Parsing

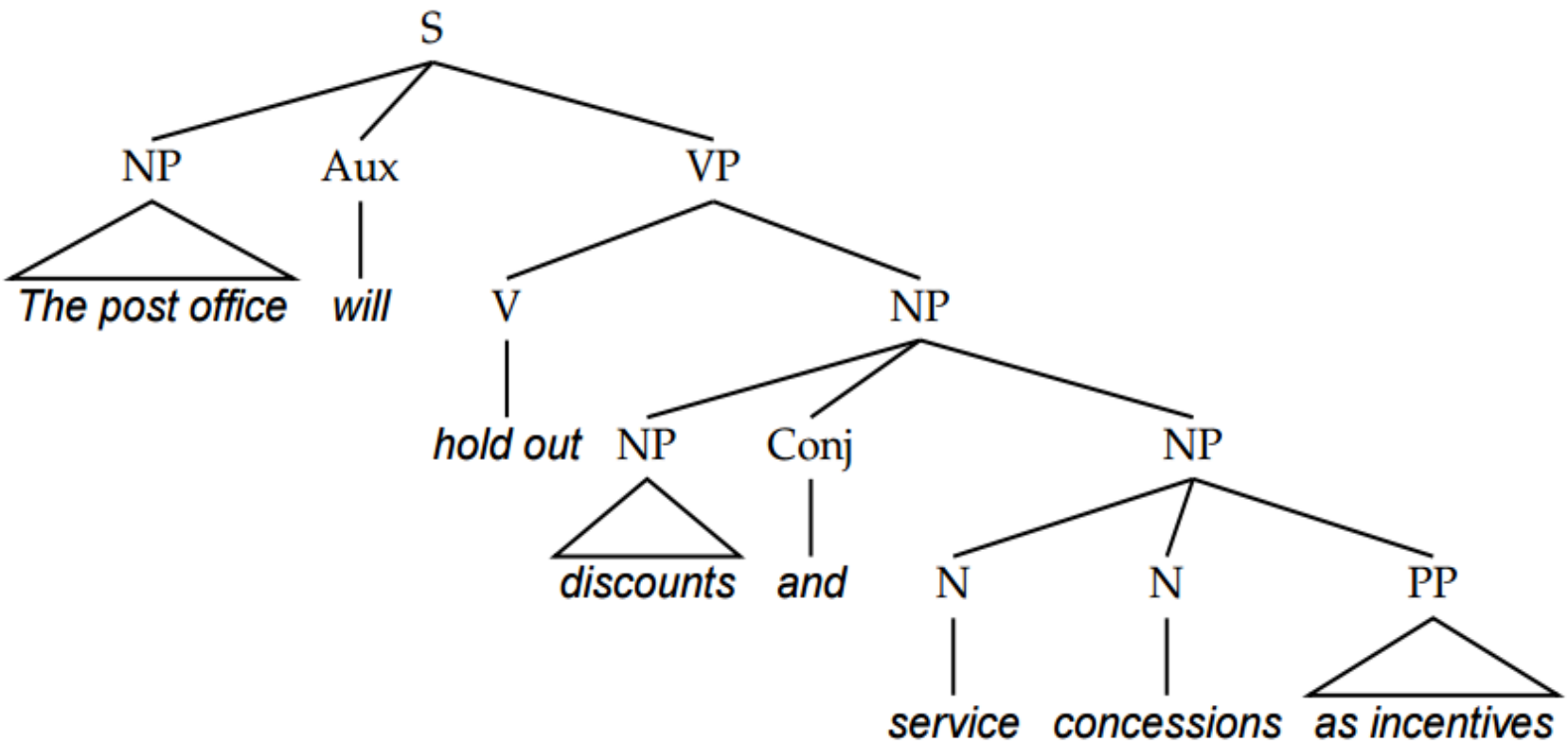
To obtain a parse tree for a sentence according to some grammar.

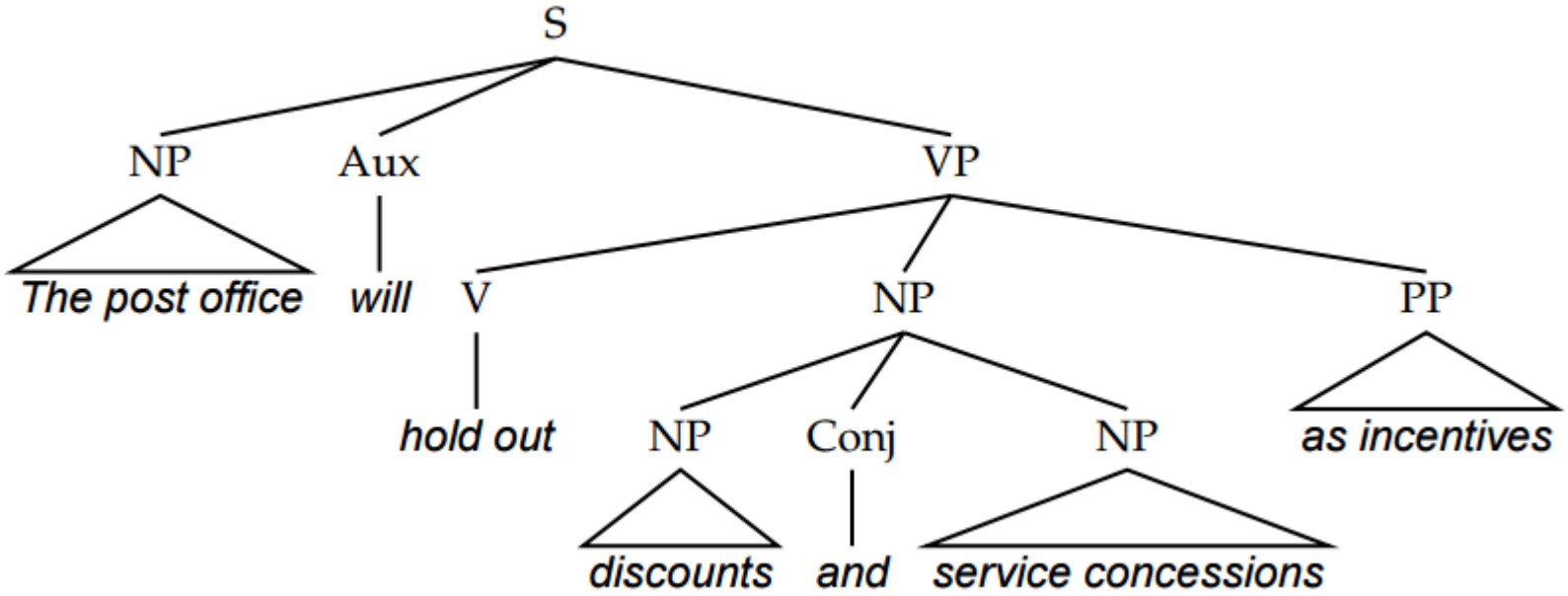


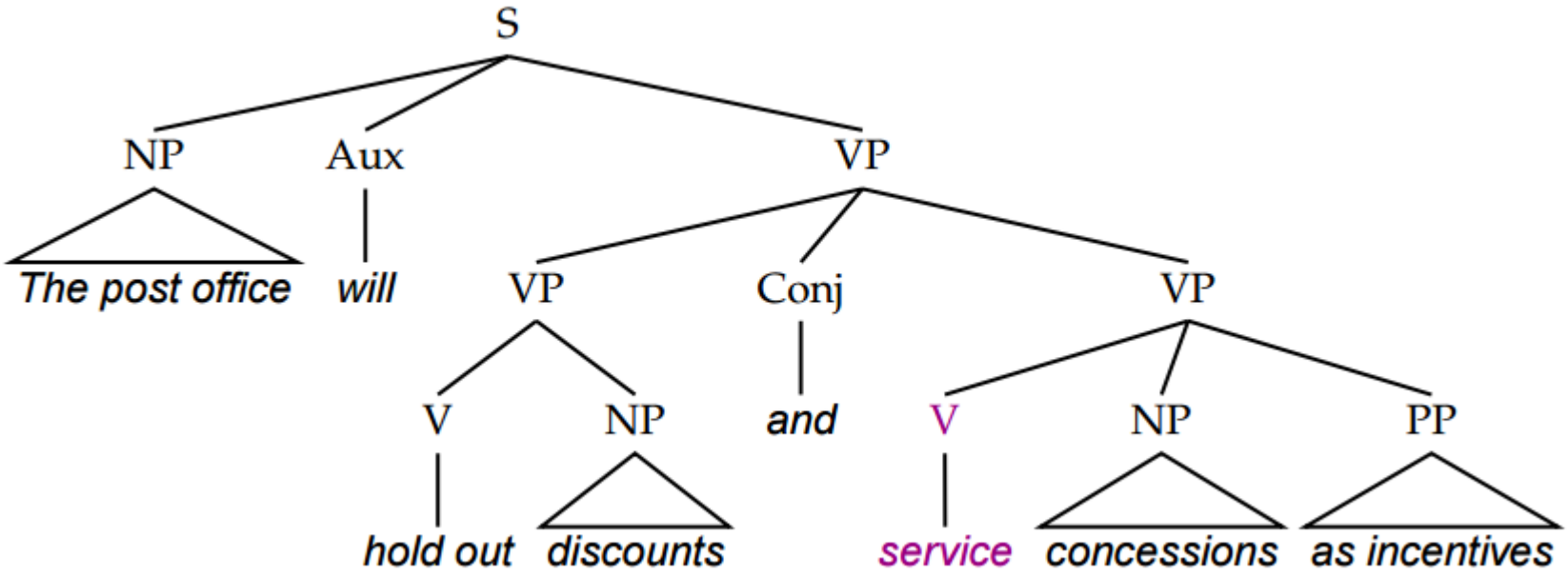
Parsing: Grammar Ambiguity

There are two or more distinct parse trees for a text.

"The post office will hold out discounts and service concessions as incentives"







“Foundations of Statistical NLP”, C. D. Manning and H. Schütze, Mit press 1999, p.410

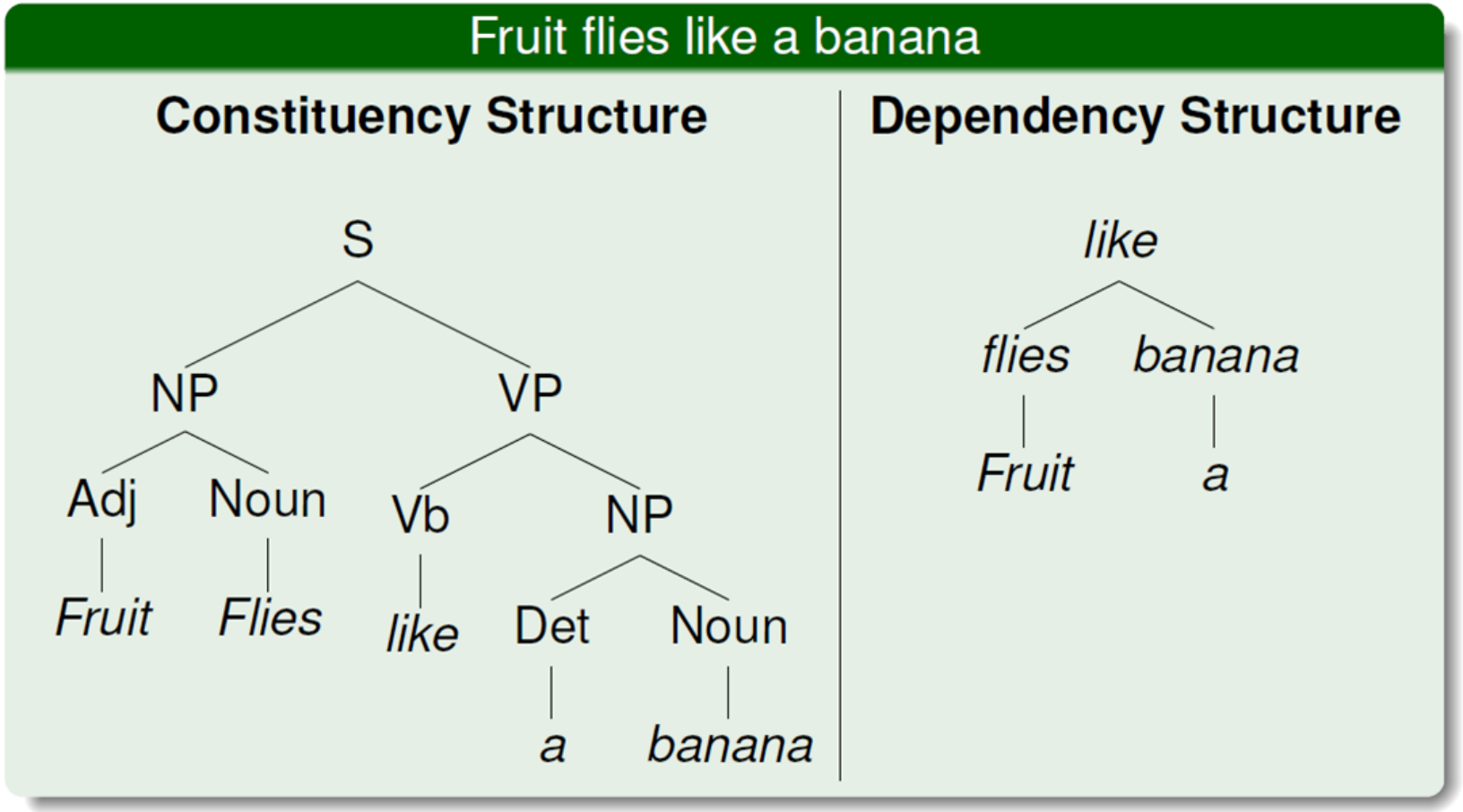
Parsing

“Were a language ever completely "grammatical" it would be a perfect engine of conceptual expression. Unfortunately, or luckily, no language is tyrannically consistent. **All grammars leak”**. Edward Sapir (Language 1921, .39)

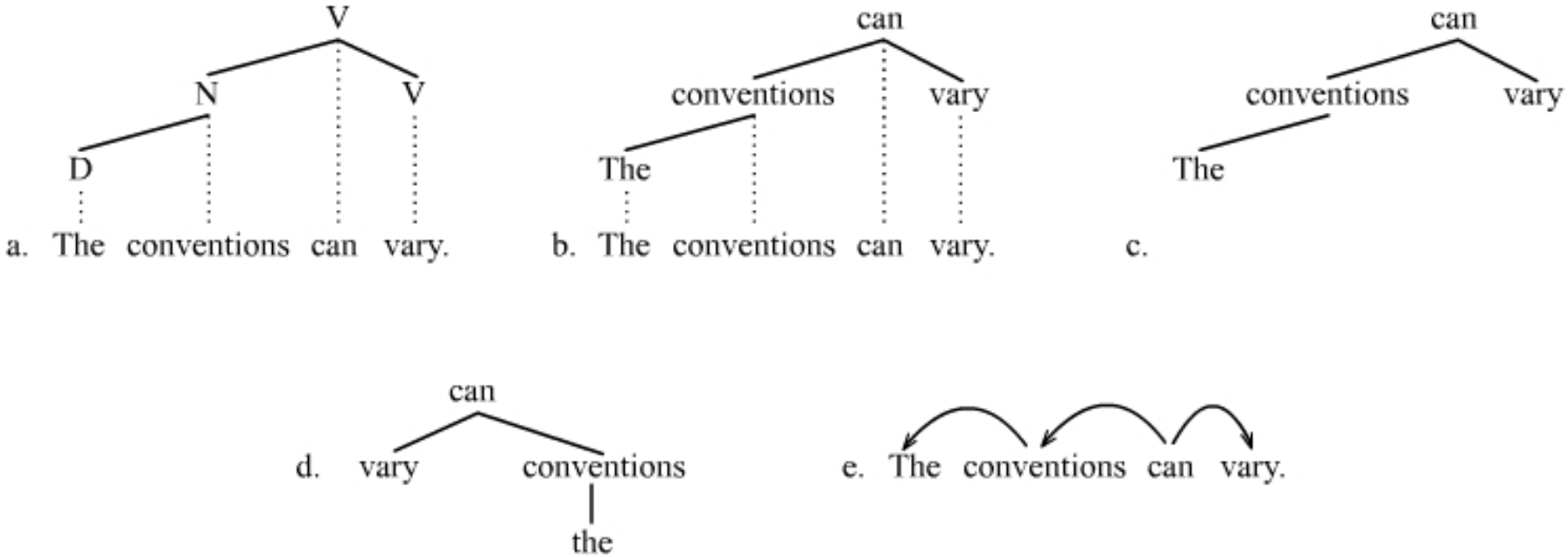
Training data (tree banks) is expensive to get but necessary for parsing training

1. Reveal hidden structure of the language
Dependency Analysis

Get word relations from a parse tree



From parse trees to word dependencies



https://en.wikipedia.org/wiki/Dependency_grammar

http://nlp.stanford.edu:8080/parser/index.jsp

Stanford Parser

The conventions can vary.

Tagging

```
The/DT  conventions/NNS  can/MD  vary/VB  ./.
```

Parse

```
(ROOT
  (S
    (NP (DT The) (NNS conventions))
    (VP (MD can)
      (VP (VB vary)))
    (. .)))
```

Universal dependencies

```
det(conventions-2, The-1)
nsubj(vary-4, conventions-2)
aux(vary-4, can-3)
root(ROOT-0, vary-4)
```

http://nlp.stanford.edu:8080/parser/index.jsp

http://www.nltk.org/book_1ed/ch08-extras.html Parsing in NLTK

1. Reveal hidden structure of the language
Semantic Role Labeling

Capturing semantic roles

SUBJ

- Dan broke [the laser pointer.]

SUBJ

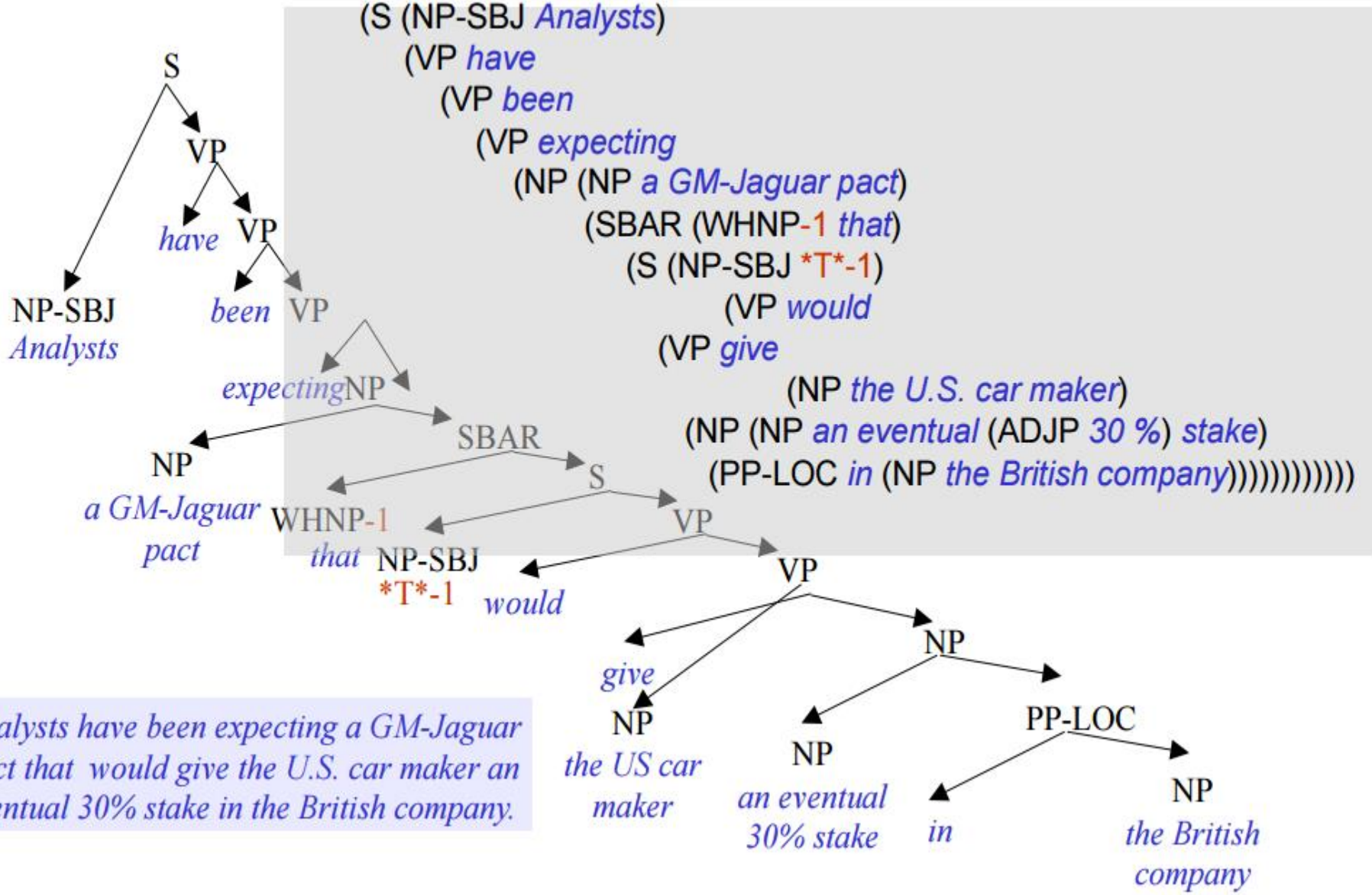
- [The windows] were broken by the hurricane.

SUBJ

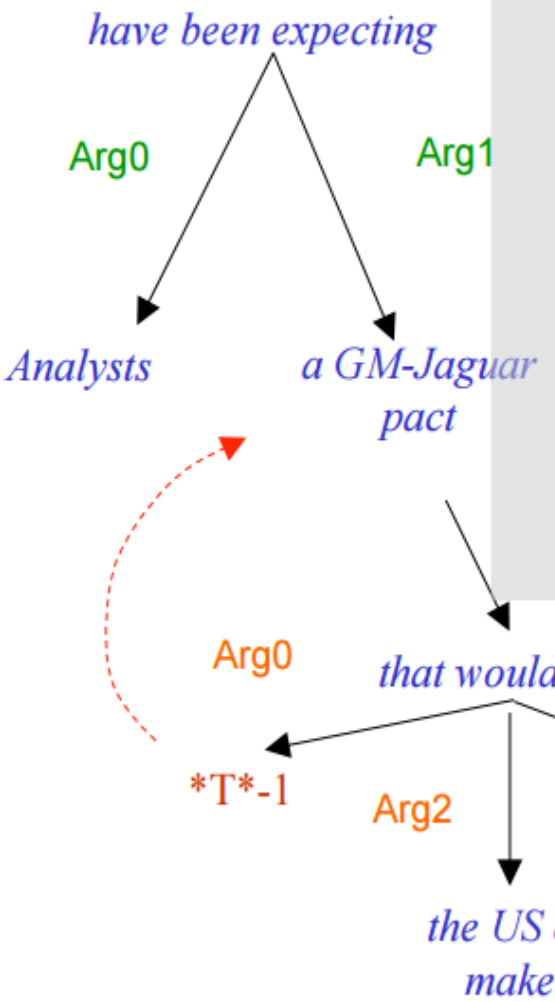
- [The vase] broke into pieces when it toppled over.

http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-1-naacl-2013-tutorial.pdf

PropBank - A TreeBanked Sentence



The same sentence, PropBanked




```
(S Arg0 (NP-SBJ Analysts)
   (VP have
      (VP been
         (VP expecting
            Arg1 (NP (NP a GM-Jaguar pact)
                  (SBAR (WHNP-1 that)
                        (S Arg0 (NP-SBJ *T*-1)
                           (VP would
                              (VP give
                                 Arg2 (NP the U.S. car maker)
                                    Arg1 (NP (NP an eventual (ADJP 30 %) stake)
                                       (PP-LOC in (NP the British
                                                  company)))))))))))
```

expect(Analysts, GM-J pact)
give(GM-J pact, US car maker, 30% stake)

Semantic Role Labeling

- PropBank Frame for *break*:

  Frameset **break.01** "break, cause to not be whole":
  Arg0: breaker
  Arg1: thing broken
  Arg2: instrument
  Arg3: pieces

- Why numbered arguments?
  - Lack of consensus concerning semantic role labels
  - Numbers correspond to verb-specific labels

PropBank seeks to assign functional tags to all modifiers or adjuncts to the verb

- Variety of ArgM's:
  - TMP - when? *yesterday, 5pm on Saturday, recently*
  - LOC - where? *in the living room, on the newspaper*
  - DIR - where to/from? *down, from Antartica*
  - MNR - how? *quickly, with much enthusiasm*
  - PRP/CAU -why? *because … , so that …*
  - REC - himself, themselves, each other
  - GOL - end point of motion, transfer verbs? *To the floor, to Judy*
  - ADV - hodge-podge, miscellaneous, "nothing-fits!"
  - PRD - this argument refers to or modifies another: *…ate the meat raw*

https://verbs.colorado.edu/~mpalmer/projects/ace.html

Task addressed by NLP (main categories)

1. Reveal hidden structure of the language, e.g. POS tagging, parsing, stemming
2. Perform “atomic” tasks of “understanding” in text, e.g. WSD, paraphrase detection, TE
3. Perform simple but tedious task involving large amount of texts, e.g. IR, QA, IE, summarization, etc.

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

“Atomic” tasks of “understanding”

- Named entity recognition
- Time and date resolution
- Negation detection and scope
- Figurative language detection
- Sarcasm detection 2016 survey at http://arxiv.org/abs/1602.03426
- Sentiment analysis
- Image/video description
- Text alignment

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

Word Sense Disambiguation

To determine the correct sense of a word in a sentence given a senses inventory, e.g. WordNet.

Senses of “bank”

```
[Synset('bank.n.01'),
Synset('depository_financial_institution.n.01'),
Synset('bank.n.03'),
Synset('bank.n.04'),
Synset('bank.n.05'),
Synset('bank.n.06'),
Synset('bank.n.07'),
Synset('savings_bank.n.02'),
Synset('bank.n.09'),
Synset('bank.n.10'),
Synset('bank.v.01'),
Synset('bank.v.02'),
… more
```

http://www.nltk.org/howto/wordnet.html Wordnet interface in NLTK

Word Sense Disambiguation

```
$ git clone https://github.com/alvations/pywsd.git
$ cd pywsd
$ ls
pywsd  README.md  test_wsd.py
$ python
>>> from pywsd.lesk import simple_lesk
>>> sent = 'I went to the bank to deposit my money'
>>> ambiguous = 'bank'
>>> answer = simple_lesk(sent, ambiguous, pos='n')
>>> print answer
Synset('bank.n.09')
>>> print answer.definition()
u'a building in which the business of banking transacted
```

https://github.com/alvations/pywsd WSD in Python

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

Synonymy/Lexical similarity

- Synonymy: synsets from WordNet or any list of synonymy pairs. E.g. “car”, “auto”, “automobile”, “motorcar”, etc.

- Lexical similarity: to determine a graded level of similarity or relatedness of a pair of words

Synonymy/Lexical similarity

Datasets with human judgments

```
midday  noon    3.94
gem     jewel   3.94
automobile      car     3.92
cemetery        graveyard       3.88
cushion pillow  3.84
boy     lad     3.82
cock    rooster 3.68
```

```
furnace implement       1.37
coast   hill    1.26
bird    woodland        1.24
shore   voyage  1.22
cemetery        woodland        1.18
```

https://github.com/DiplomadoACL/problemasenclase/tree/master/lexical_similarity/en

Task addressed by NLP (main categories)

1. Reveal hidden structure of the language, e.g. POS tagging, parsing, stemming
2. Perform “atomic” tasks of “understanding” in text, e.g. WSD, paraphrase detection, TE
3. Perform simple but tedious task involving large amount of texts, e.g. IR, QA, IE, summarization, etc.

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

Paraphrase detection

given a pair of sentences, classify them as paraphrases or not paraphrases

- **Sentence 1**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
- **Sentence 2**: Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- **Class**: 1 (true paraphrase)

http://aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_(State_of_the_art)

Semantic Textual Similarity (STS)

The degree of semantic equivalence between two sentences

- **100%** “The bird is bathing in the sink.” “Birdie is washing itself in the water basin.”
- **80%** “In May 2010, the troops attempted to invade Kabul.” “The US army invaded Kabul on May 7th last year, 2010.”
- **60**% “John said he is considered a witness but not a suspect.” "He is not a suspect anymore. John said.”

Semantic Textual Similarity (STS)

The degree of semantic equivalence between two sentences

- **40%** “They flew out of the nest in groups. They flew into the nest together.”
- **20%** “The woman is playing the violin.” “The young lady enjoys listening to the guitar.”
- **0%** “John went horse back riding at dawn with a whole group of friends.” “Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.”

https://www.cs.york.ac.uk/semeval-2012/task6/index.html

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

Textual Entailment

Is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. In the TE framework, the entailing and entailed texts are termed *text* and *hypothesis*.

The entailment need not be pure logical - it has a more relaxed definition: "t entails h (t ⇒ h) if, typically, a human reading t would infer that h is most likely true."[

http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal

Textual Entailment

- text entails hypothesis
  - T: *If you help the needy, God will reward you*.
  - H: *Giving money to a poor man has good consequences*.
- text contradicts hypothesis
  - T: *If you help the needy, God will reward you*.
  - H: *Giving money to a poor man has no consequences*.
- text does not entail nor contradict
  - T: *If you help the needy, God will reward you*.
  - H: *Giving money to a poor man will make you better person*.

http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal

“Atomic” tasks of “understanding”

- Word sense disambiguation
- Synonymy/Lexical similarity
- Paraphrase detection/Textual similarity
- Textual Entailment
- Anaphora resolution/coreference
- Collocation detection
- Multiword expressions detection

Introduction to Anaphora

- Anaphora in Etymology
  - Ancient Greek : Anaphora = αναφορα (Anajora)
    - ανα (Ana) → *back in an upward direction*
    - φορα *(Jora)* → *the act of carrying back upstream*
- Example:
  - **The Empress** hasn't arrived yet but **she** should be here any minute.
    - she →Anaphor
    - The Empress (NP) → Antecedent
    - Empress (N) is NOT the antecedent !
    - **Coreferent** → Both The Empress and she refer to the same REAL WORLD ENTITY

http://www.slideshare.net/kancho/anaphora-resolution

Catafora

- Catafora
  - when the “anaphor” precedes the “antecedent”
  - Because **she** was going to the post office, **Julie** was asked to post a small parcel

http://www.slideshare.net/kancho/anaphora-resolution

Anaphora Resolution (AR)

- Anaphora Resolution(AR) is the process of determining the antecedent of an anaphor.
  - **Anaphor** – The reference that points to the previous item
  - **Antecedent** –The entity to which anaphor refers
- Needed to derive the "Correct Interpretation" of a text
- Is a complicated problem in NLP !

http://www.slideshare.net/kancho/anaphora-resolution

Coreferential chain

- when the anaphor and more than one of the preceding (or following) entities (usually noun phrases) have the same referent and are therefore pairwise coreferential

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.

- *she* ⇒ *Sophia Loren*
- *the actress* ⇒ *Sophia Loren*
- *the U2 singer* ⇒ *Bono*
- *her* ⇒ *Sophia Loren*
- *she* ⇒ *Sophia Loren*

Coreference Chains:

- {Sophia Loren, she, the actress, her, she}
- {Bono, the U2 singer }
- {a thunderstorm}
- {a plane}

http://www.slideshare.net/kancho/anaphora-resolution

Tedious tasks involving large amount of texts

- Information retrieval
- Question answering (factoids, inference)
- Information extraction (fact, relation extraction)
- Text classification and clustering
- Translation
- Summarization

Tedious tasks involving large amount of texts

- Dialog systems (synthesis/analysis)
- Student short-answer grading
- Topic detection
- Ontology learning
- Lexical and structural simplification