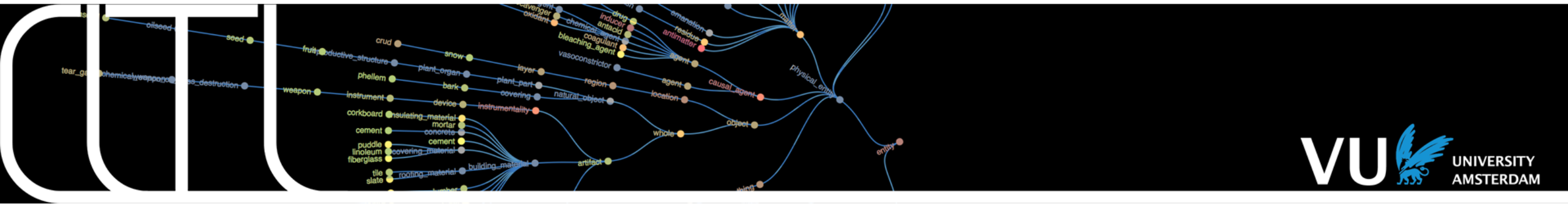
Text Mining CBS 2019

Lecture 1: Linguistics and Natural Language Processing

Piek Vossen

kyoto.let.vu.nl/~vossen/cbs/cbs-lecture-1-linguistics-nlp.pdf



Logistiek

- 9:30 - 10:30 lecture
- pauze
- 11:00 - 12:00 lecture
- lunch
- 13:00 - 14:00 lab sessie
- pauze
- 14:30 - 15:30 lab sessie

- eigen laptop met voldoende schijfruimte
- downloaden en installeren:
  - Anaconda (Python 3.7): https://anaconda.org
  - https://github.com/cltl/text-mining-ba
- Jupiter notebooks: https://jupyter.org
- NLTK:
- - http://www.nltk.org/book
- SpaCy
- - https://spacy.io
- - https://spacy.io/usage/models
- Skitlearn:
- - https://scikit-learn.org/stable/

Welke informatie staat er (niet) in een tekst

- Michael van Gerwen is er zondag net niet in geslaagd om de Dutch Darts Masters in Zwolle op zijn naam te schrijven. De Brabander verloor in een bloedstollende finale met 7-8 van de Engelsman Ian White.

- De dertigjarige Van Gerwen verloor in de eindstrijd meteen zijn eigen leg en kwam met 0-2 achter. Hij knokte zich vervolgens terug en maakte de break achterstand in de zevende leg ongedaan met een finish van 170.

- De 48-jarige White raakte hier niet van onder de indruk en kwam met 6-4 voor. 'Dab Diamond' gaf zijn voorsprong opnieuw uit handen, maar Van Gerwen slaagde er niet in om de partij te beslissen.

https://www.nu.nl/darts/5909316/van-gerwen-grijpt-net-naast-zesde-titel-op-rij-bij-dutch-darts-masters.html

Welke informatie staat er

Entities:
- Michael van Gerwen, De dertigjarige Van Gerwen, De Brabander, Hij, Van Gerwen
- de Engelsman Ian White, De 48-jarige White, ‘Dab Diamond’

- **Michael van Gerwen** is er zondag net niet in geslaagd om de **Dutch Darts Masters** in **Zwolle** op zijn naam te schrijven. **De Brabander** verloor in een bloedstollende finale met 7-8 van **de Engelsman Ian White**.

- **De dertigjarige Van Gerwen** verloor in de eindstrijd meteen zijn eigen leg en kwam met 0-2 achter. **Hij** knokte zich vervolgens terug en maakte de break achterstand in de zevende leg ongedaan met een finish van 170.

- **De 48-jarige White** raakte hier niet van onder de indruk en kwam met 6-4 voor. **'Dab Diamond'** gaf zijn voorsprong opnieuw uit handen, maar **Van Gerwen** slaagde er niet in om de partij te beslissen.

Welke informatie staat er

Events
geslaagd, op zijn naam schrijven, verliezen, finale, verloor, eindstrijd, kwam achter, knokte terug, maakte ongedaan, raakte onder de indruk, kwam voor, gaf uit handen, slaagde, partij beslissen

- **Michael van Gerwen** is er zondag net niet in **geslaagd** om de **Dutch Darts Masters** in **Zwolle** **op zijn naam te schrijven**. **De Brabander** **verloor** in een bloedstollende **finale** met 7-8 van **de Engelsman Ian White**.

- **De dertigjarige Van Gerwen** **verloor** in de **eindstrijd** meteen zijn eigen leg en **kwam** met 0-2 **achter**. Hij **knokte** zich vervolgens **terug** en **maakte** de break achterstand in de zevende leg **ongedaan** met een finish van 170.

- **De 48-jarige White** **raakte** hier niet van **onder de indruk** en **kwam** met 6-4 **voor**. **'Dab Diamond'** **gaf** zijn voorsprong opnieuw **uit handen**, maar **Van Gerwen** **slaagde** er niet in om de **partij** te **beslissen**.

Welke informatie staat er

Judgements: bloedstollende
Time: zondag, meteen, vervolgens, opnieuw
Factuality: net niet hier
Collocational: er in slagen
Domain: leg, break

- **Michael van Gerwen is er zondag net niet** in **geslaagd** om de **Dutch Darts Masters** in **Zwolle op zijn naam te schrijven**. **De Brabander verloor** in een **bloedstollende finale** met 7-8 van **de Engelsman Ian White**.

- **De dertigjarige Van Gerwen verloor** in de **eindstrijd meteen** zijn eigen leg en **kwam** met 0-2 **achter**. Hij **knokte** zich **vervolgens terug** en **maakte** de break achterstand in de zevende leg **ongedaan** met een finish van 170.

- **De 48-jarige White raakte hier niet** van **onder de indruk** en **kwam** met 6-4 **voor**. **'Dab Diamond' gaf** zijn voorsprong **opnieuw uit handen**, maar **Van Gerwen slaagde er niet in** om de **partij** te **beslissen**.

Maar je hebt linguïstiek nodig

- **De 48-jarige White raakte hier niet** van **onder de indruk** en **kwam** met 6-4 **voor**. **'Dab Diamond' gaf** zijn voorsprong **opnieuw uit handen**, maar **Van Gerwen slaagde er niet in** om de **partij** te **beslissen**.

- Coreference: Dab Diamond = White, zijn == Dab Diamons
- Down casing: De -> de, Van Gerwen -> Van Gerwen
- indruk: zelfstandig naamwoord of werkwoord?
- uit handen geven: onderwerp, lijdend voorwerp
- “er niet in om”

CBS Text Mining cursus

Datum	Theorie	Lab sessies	
3 - 4 juni	Piek Vossen • Kennis van taal • Regels, lexica, machine learning, data annotatie • Sentiment, opinies, en emoties	Filip Ilievski & Piek Vossen • Python toolkits NLTK en SpaCy • Sentiment: lexicon en statistiek • Evaluatie: accuracy, recall, precision, f1	Basis
13-14 juni	Piek Vossen • Van Tekst naar Feature representatie • Entiteiten • Eigenschappen van entiteiten	Filip Ilievski & Piek Vossen • Entiteiten detectie en classificatie • Entiteiten disambiguering/linking • Extractie van eigenschappen	
1-2 juli	Antske Fokkens • Word embeddings • Neurale netwerken • Blackbox versus Clearbox	Pia Sommerauer & Antske Fokkens • Similarity and Relatedness • Embeddings in Machine Learning	Verdieping

Overview of lecture

- Part I: Linguistics and Natural Language Processing (NLP)
    - morphology
    - syntax
    - semantics
- Part II: NLP Pipelines
- *Part III: Language as data*

Part I: Linguistics

- Language and Structure
- Language and Meaning
- *Language as data*

Linguistics

Subdiscipline	Medium or unit	Natural language module
phonetics, phonology	sounds	Automatic Speech Recognition
morphology	words, word formation	Part-Of-Speech taggers, lemmatisers
syntax	sentences, grammatical structure and function	Syntactic parsers, chunkers
semantics	meaning	Semantic parsers
pragmatics	language use in context	Context and domain models
methods	introspection, behavorism, neuro-cognitive models, empirical (experimental & stochastic), mathematical models	
resources	Lexicons, grammars, data collections and annotations, data models, annotations	

Morphology

- Study of form and structure of words
- Words are composed of ***morphemes***
- ***Morpheme*** is the smallest meaning-bearing unit:
  - e.g. *talked* contains two morphemes: *talk* and *-ed* (past)

Types of morphemes

- **Free Morphemes**: occur independently, e.g. *boy*, *sing*

- **Bound Morphemes**: attached to another morpheme, and cannot be used independently, e.g.
    - English [NUMBER pl] -*s* → boys,
    - Dutch [NUMBER pl] -*s* → appels, [NUMBER pl] -*en* → appelen

- **Affixes**: **prefixes** (e.g. *ge*lopen), **infixes** (e.g. burgemeester*s*post, **suffixes** (e.g. loop*je*)

Some other basic terms

- **Root**: an unanalysable morpheme, expressing the basic lexical content of a word. Also defined as ‘what is left of a complex form when all affixes are stripped’.

- **Stem**: consists of at least a root. It can contain (a) derivational affix(es).

- **Base**: a morpheme to which an affix may be added. A base may be simplex (root) or complex (root + affixes).

Part of Speech (PoS)

- Words have **part-of-speech** (**PoS**), which specifies the typical phrase structures in which they can be the head (see later)

- **Open class** (open to word formation and neologisms):

    - Noun (N, *boat*), Verb (V, *float*), Adjective (A, *large, fast*), Adverb (*very, largely*)

        - new words invented every day and other words are forgotten, e.g. "***belubberen***";

        - millions of open class words if we include specialised language (chemistry, medicine, product names)

- **Closed class** (you can not invent a new closed class word):

    - Pronoun (PRN, *he, him, this, who*), Proposition (P, *in, at, from, in front of*), etc.

    - relatively fixed, slowly change over generations; small set of less than a hundred words

Word modification

- Given a root, base or stem derive different forms
    - **Inflection**, expresses syntactic properties such as person (1,2,3), number (singular, plural), gender, tense, e.g. “book*s*”, “*her* hair”, “walk*ed*”.
    - **Derivation**: changes semantic and grammatical properties, e.g. “*in*appli*cable”* A, “head” N —> “*be*head” V
        - **Compounding**: “beach head”, “tarwemeel” (oat flour), “kindermeel” (child flour)
        - **Combinations**: aircraft-carriers = ((air+craft)+(carry - er)) (not: air +craftcarrier …)
- Word formation is very productive, our lexicon is potentially infinite:
    - the number of unseen compounds detected in German & Dutch newspapers grows linearly with the number of newspapers over time
    - the names for new chemical compounds and proteins grow rapidly every year
    - new products (e.g. apps) launched every year

Inflectional Morphology

- Inflection is required by syntactic criteria, e.g. an English verb must have tense, a noun is singular or plural

- It marks grammatical (=morpho-syntactic) distinctions:
  - Conjugation (verbal categories): person (1, 2, 3), number, gender
    - tense, aspect, mood, agreement
  - Declination (nominal categories)
    - case, number (singular, plural), gender, degree, definiteness

Derivational changes

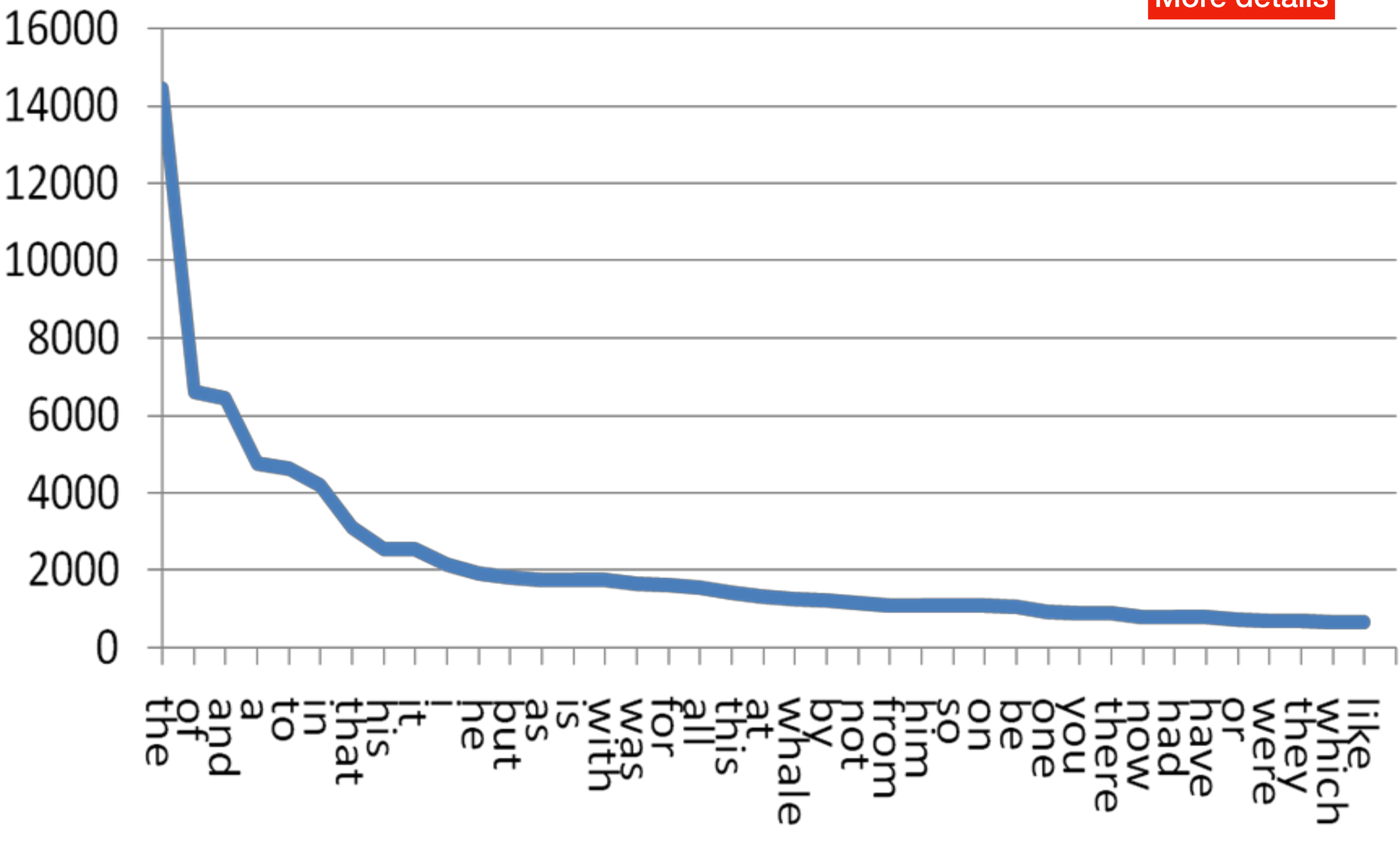
- **semantics**,
  - e.g. [clear] → [un+[clear]] = unclear
- **syntactic category**,
  - e.g. [derive]V → [[[derive]V +ation]N +al]Adj = derivational
- **valency** of a verb,
  - e.g. [qaw] ’it breaks’ → [t+[qaw]] ’he breaks it’ (Havasupai Indian language)
- several from the above,
  - e.g. [understand]V → [[understand]V +able] = understandable

Forms in language

- A language has:
    - 11-112 phonemes (sound units)
    - 4,000-10,000 morphemes (word units)
    - 50,000 common words, millions of words including terminologies
    - An infinite number of sentences
- But we use small proportion of these forms very frequently even though we recognise and understand most of those

Power law distribution of word frequency

More details



More details

Zipfian distributions

$f(w_i)=f(W_1)/r_i(w_i)$

- The frequency of a word in a ranked list is the equal to the frequency of the most frequent word divided by the rank 100, 50, 30, 25, 20, 16, 14, etc….

- Most frequent words also tend to be short and have many different meanings

George Kingsley Zipf
1902 - 1950

Lexicon of forms

- Lists all common base forms (a hundred of thousand in a standard dictionary) with:
  - their part-of-speech
  - inflectional paradigm
  - typical (conventional) derived forms
- Inflectional paradigms and derivational morphemes

Morphology in Computational Linguistics

- Analysing **complex words**, defining their component parts: anti+dis+establish+ment+arian+ism

- Analysis of **grammatical information**, encoded in words:

    - *sings*

        - **Part-of-speech** = VERB

        - **Inflectional information** = [PERSON 3, NUMBER singular,TENSE present

- Obtaining the **stem** or **root**: to reduce size of the data, to find the word in the lexicon

    - Dutch “stemmen” (voice or vote) —> “stem” noun, “stemmen” verb

    - Reduction of lexicon size (English 2:1, Dutch/German 5:1, Finnish/Turkish >200:1) (Crysmann 2006)

Part-of-Speech tagging

- Task: assign the Part-of-Speech category (e.g. noun, verb, adjective) to every token and add the lemma

- Tagset: no consensus (there are at least 50):
    - http://universaldependencies.org/u/pos/

- PoS tagging is done using machine learning
    - Hidden Markov Models, Decision Trees, SVM, Naive Bayes

- Main challenge:
    - Data sparseness for specific languages and domains

PoS-tagging

- Assign morphosyntactic categories to words in a specific context:

The	green	train	runs	down	that	track	.
Det	Adj/NN	NNS/ VBZ	NN/ VB	Prep/ Adv/	SC/ Pron	NN/ VB	.
Det	Adj	NNS	VB	Prop	Pron	NN	.

- Lexical and contextual constraints are used to identify the right tag

More details

Markov model

- Markov models are used to predict sequences
- They assume that the next state in the sequence is dependent on the current state (only)
- See Section on PoS tagging for more in-depth workings: http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/ambiguity/ambiguity-2009.html

State-of-the-art in PoS-tagging

- Accuracy around ***95-97%*** for all tokens when training and testing on the same domain

- Remaining issues:
  - long distance dependencies
  - genuine ambiguities
  - annotation errors
  - unknown words, data sparseness

PoS-tagging issues

- Better/richer models? (we would need even more data!)
- ***Coverage for other domains can drop to 75%***
- Morphologically rich languages need far larger training data sets (Finnish & Turkish need up to 10x more data than English)
- 95% sounds good but you don’t know which tokens are wrongly tagged
  - Relatively high proportion of sentences has at least one error
  - Errors propagate: wrong PoS may lead to wrong word sense, named entity, parse tree etc.

Morphology tooling

- **NLTK**: https://www.nltk.org/book/ch05.html
  - Annotated corpora with part-of-speech tags
  - Dictionaries with word forms, their segmentation, part-of-speech text and inflectional information
    - various morphological parsers: stemming and PoS annotation
- Dutch morphological lexicon **e-Lex**: https://ivdnt.org/downloads/tstc-e-lex
- Some famous stemmers, lemmatisers and taggers:
  - **Porter** stemmer: https://tartarus.org/martin/PorterStemmer/
  - **Snowball**: https://snowballstem.org
  - **Treetagger**: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

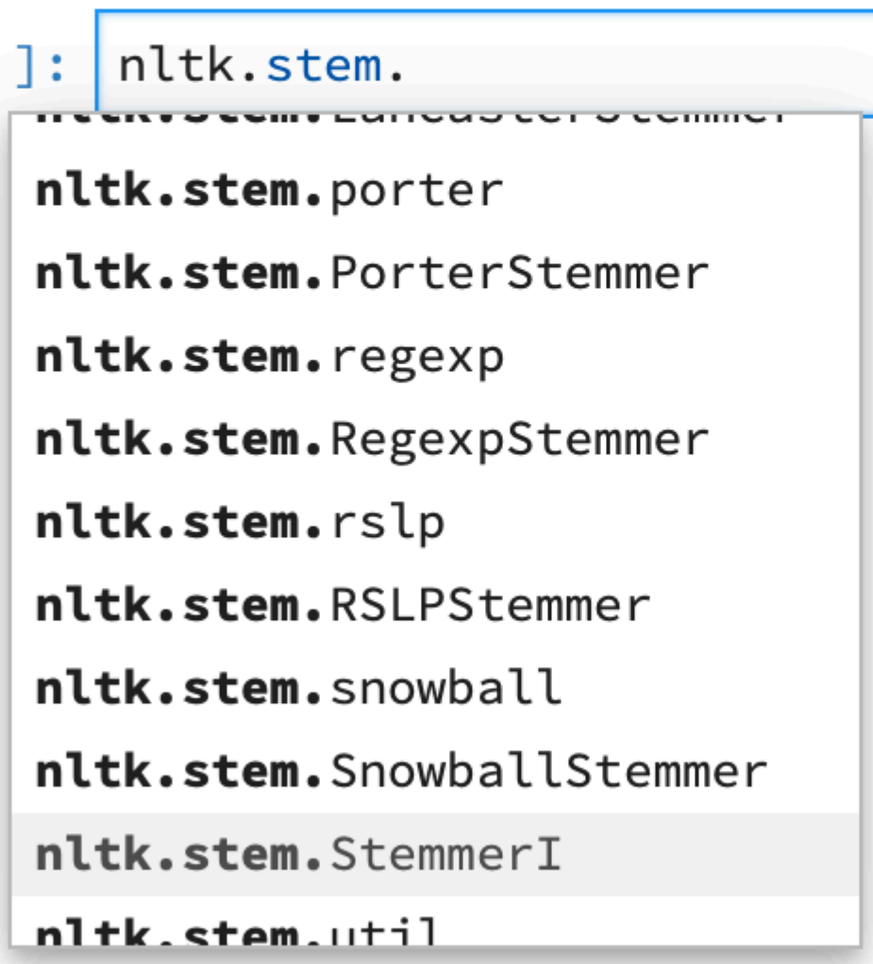
Morphology tooling in NLTK

```
In [1]: import nltk

In [2]: from __future__ import print_function

In [3]: from nltk.stem import *

In [ ]: nltk.stem.
```



Annotated texts

NLTK: nltk_data/corpora/brown/ca11

- The/at Birds/nns-tl got/vbd five/cd hits/nns and/cc all/abn three/cd of/in their/pp$ runs/nns off/in Kunkel/np before/cs Hartman/np took/vbd over/rp in/in the/at top/nn of/in the/at fourth/od ./.

- Hartman/np ,/, purchased/vbn by/in the/at A's/nn from/in the/at Milwaukee/np Braves/nns-tl last/ap fall/nn ,/, allowed/vbd no/at hits/nns in/in his/pp$ scoreless/jj three-inning/jj appearance/nn ,/, and/cc merited/vbd the/at triumph/nn ./.

- Keegan/np ,/, a/at 6-foot-3-inch/jj 158-pounder/nn ,/, gave/vbd up/rp the/at Orioles'/nps$ last/ap two/cd safeties/nns over/in the/at final/jj three/cd frames/nns ,/, escaping/vbg a/at load/nn of/in trouble/nn in/in the/at ninth/od when/wrb the/at Birds/nns-tl threatened/vbd but/cc failed/vbd to/to tally/vb ./.

Morphological lexicon

E-Lex: https://ivdnt.org/downloads/taalmaterialen/tstc-e-lex

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[DET:<de>][HD:<stem>]\\

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[DET:<de>][HD:<stem>][PC:[PP:[HD:<op>][OBJ1:NP]]]\\

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[DET:<de>][HD:<stem>][PC:[PP:[HD:<tegen>][OBJ1:NP]]]\\

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[DET:<de>][HD:<stem>][PC:[PP:[HD:<voor>][OBJ1:NP]]]\\

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[DET:<een>][HD:<stem>]\\

97809\stem\(stem)[N]\\A0;C->N->S4;C->N->Z2\A0;S4;Z2\\295418\stem\N(soort,ev,basis,zijd,stan)\\V\stEm\stEm\stEm\'stEm\V\686\[HD:<stem>]\\

…..

97810\stemmen\(stem)[V]\\\\\471636\gestemd\WW(vd,prenom,zonder)\B\C\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\0\\\

97810\stemmen\(stem)[V]\\\\\107427\gestemd\WW(vd,vrij,zonder)\B\V\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\108\[SU:NP][HD:<hebben>][VC:[HD:<gestemd>]]\\

97810\stemmen\(stem)[V]\\\\\107427\gestemd\WW(vd,vrij,zonder)\B\V\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\108\[SU:NP][HD:<hebben>][VC:[PPART:[HD:<gestemd>]]]\\

97810\stemmen\(stem)[V]\\\\\107427\gestemd\WW(vd,vrij,zonder)\B\V\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\108\[SU:NP][HD:<hebben>][VC:[PPART:[OBJ1:NP][HD:<gestemd>]]]\\

97810\stemmen\(stem)[V]\\\\\107427\gestemd\WW(vd,vrij,zonder)\B\V\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\108\[SU:NP][HD:<hebben>][VC:[PPART:[OBJ1:NP][PC:[PP:[HD:<tot>][OBJ1:INF]]][HD:<gestemd>]]]\\

97810\stemmen\(stem)[V]\\\\\107427\gestemd\WW(vd,vrij,zonder)\B\V\x@stEmt\G@stEmt\x@stEmt\x@-'stEmt\V\108\[SU:NP][HD:<hebben>][VC:[PPART:[OBJ1:NP][PC:[PP:[HD:<tot>][OBJ1:NP]]][HD:<gestemd>]]]\\

…..

Multiword expressions

- Fixed Idioms
  - *An apple a day keeps the doctor away*
  - *kick the bucket, Raining cats and dogs*
- Less fixed idioms
  - *shooting from the hip*
- Slots
  - *X, let alone Y*
- Collocations
  - *the engine is running, strong coffee, count on, treat for*
- Selection restrictions:
  - *essen/fressen, a glass of ...*

Syntax

- We experience a sentences as a complete grammatical structure.
- We can freely combine words into **phrases** or **constituents** and we have a strong intuition about the grammaticality of these structures within a sentence.
- What is a **phrase** or **constituent**?
  - A phrase is a word or a group of words which functions as single unit within a grammatical hierarchy
  - A phrase is built around a **head** lexical item and has a certain syntactic behaviour
    - she ⇒ Noun Phrase or NP (the **head** is a pronoun)
    - a very beautiful morning ⇒ NP (the **head** is a Noun)
    - chases the cat ⇒ Verb Phrase or VP (**head** is a Verb)

Syntactic elements

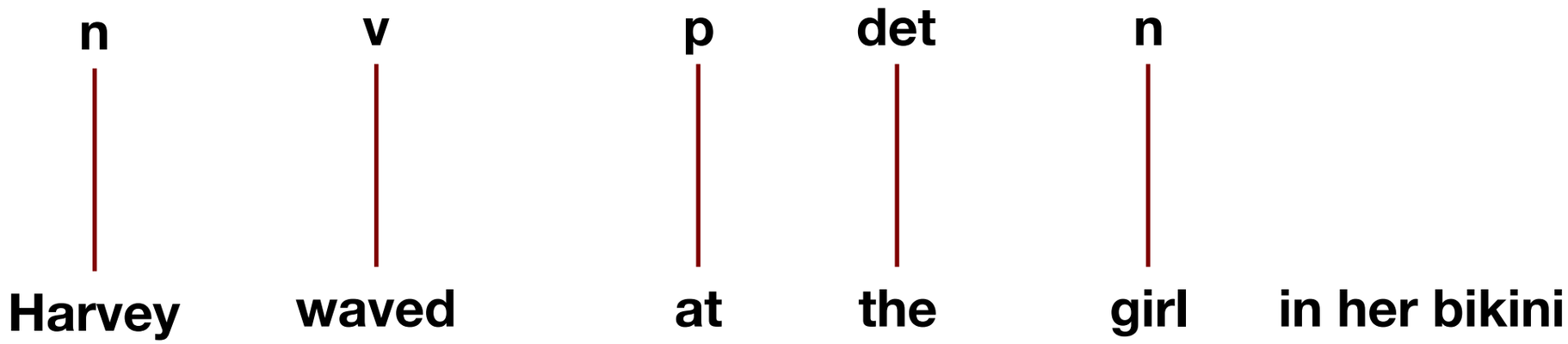
- Phrasal categories: Noun Phrase (NP)
    - Prepositional Phase (PP)
    - Verb Phrase (VP)
    - Adverbial Phrase (AdvP)
    - Adjectival Phrase (AP)
- Lexical categories: Noun (N)
    - Pronoun (Pr)
    - Adjective (A)
    - Adverb (Adv)
    - Verb (V)
    - Preposition (P)

Syntax

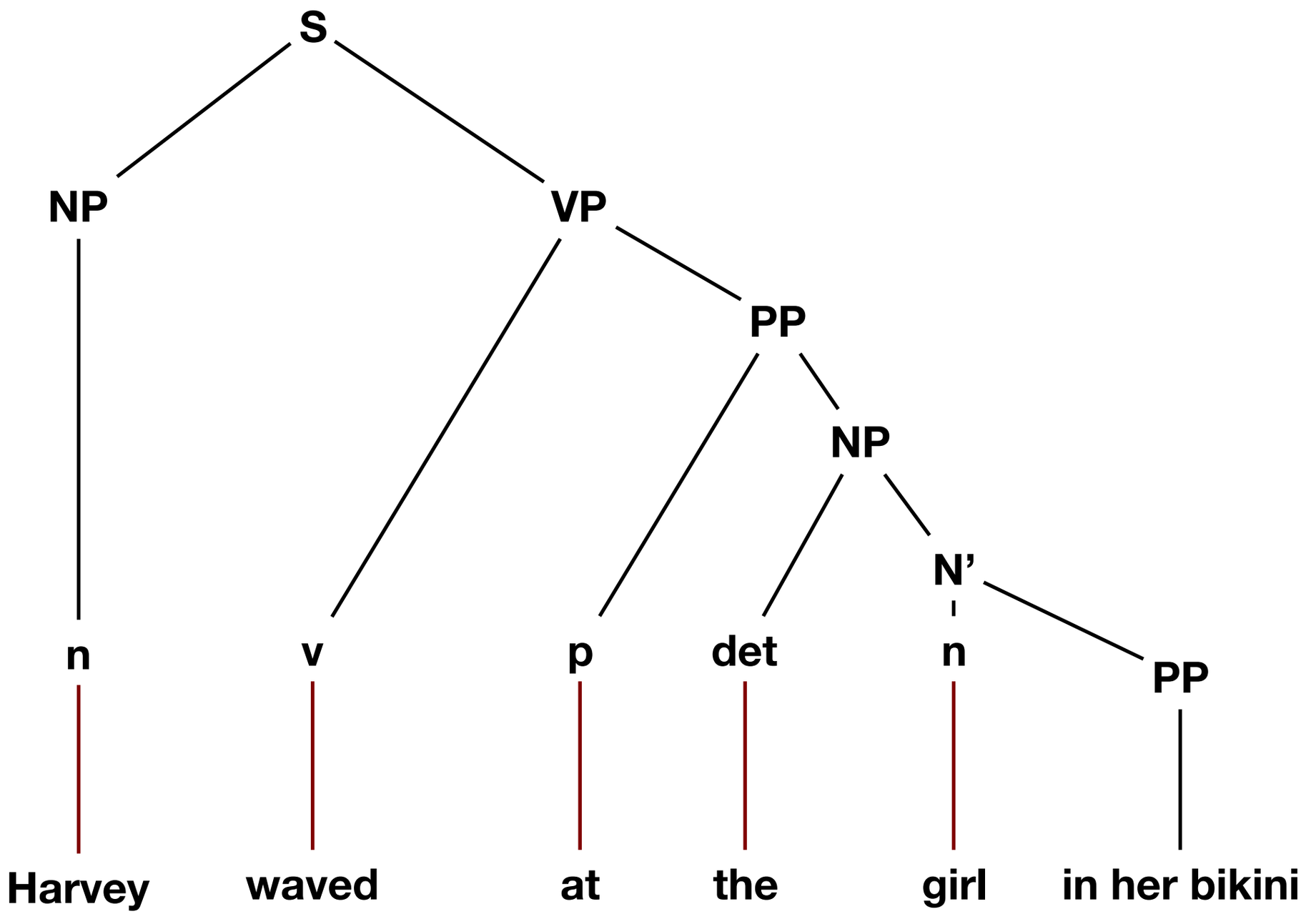
- Phrases can be nested hierarchically:
  - very nice = Ajective Phrase or AP (head is an adjective (A))
  - a very nice looping = NP (head is noun (N)
    - performs a nice looping = VP (head is a verb (V))
    - with a long stick = Prepositional Phrase or PP (head is preposition (P))
    - the cow performs a very nice looping with a long stick = Sentence (S)
- Functions
- Dependency relations between the heads of the constituents: subject (cow), object (looping), main verb (perform), modifier (nice), adjunct (stick)

Syntactic trees

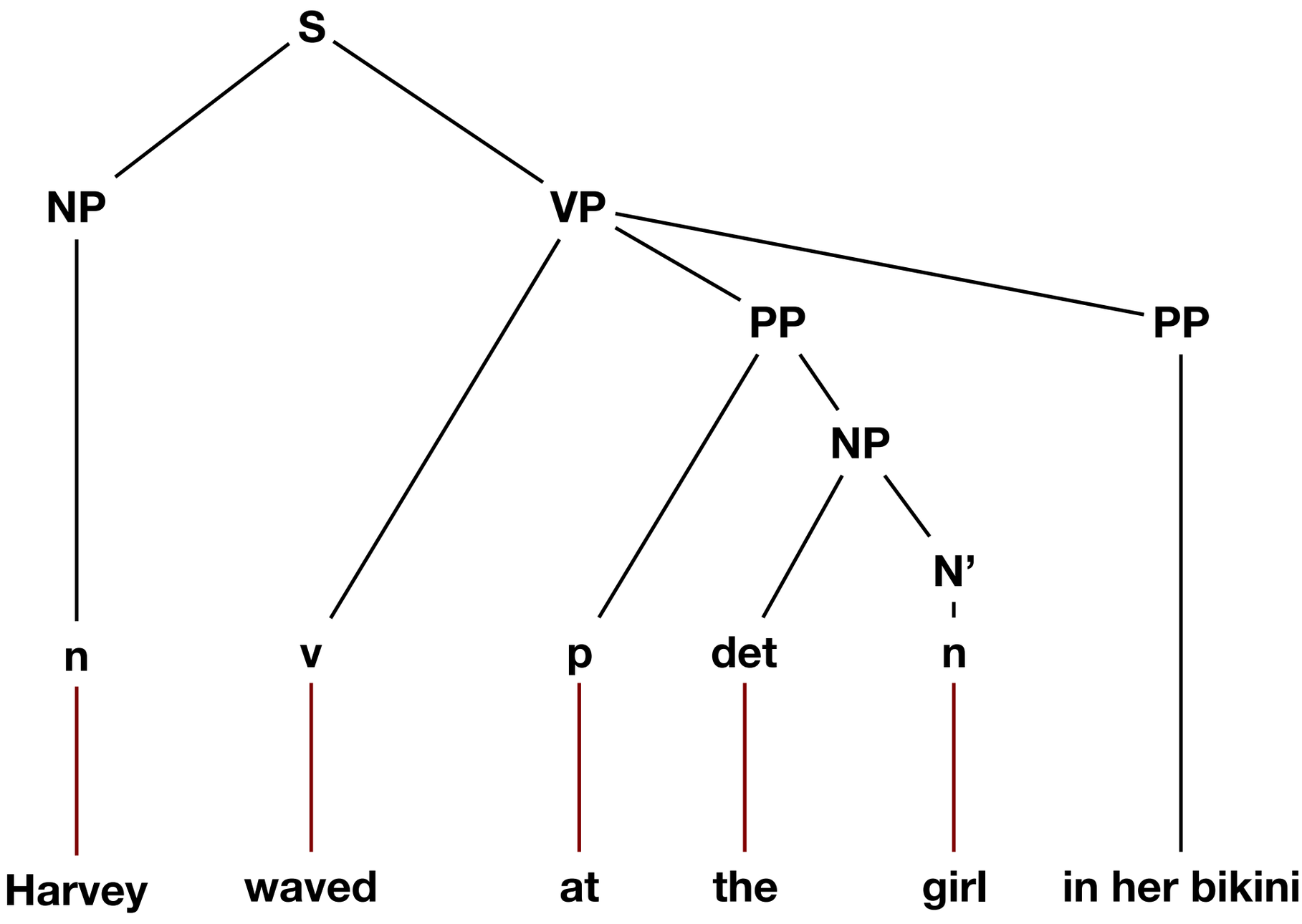
Syntactic trees



Syntactic trees



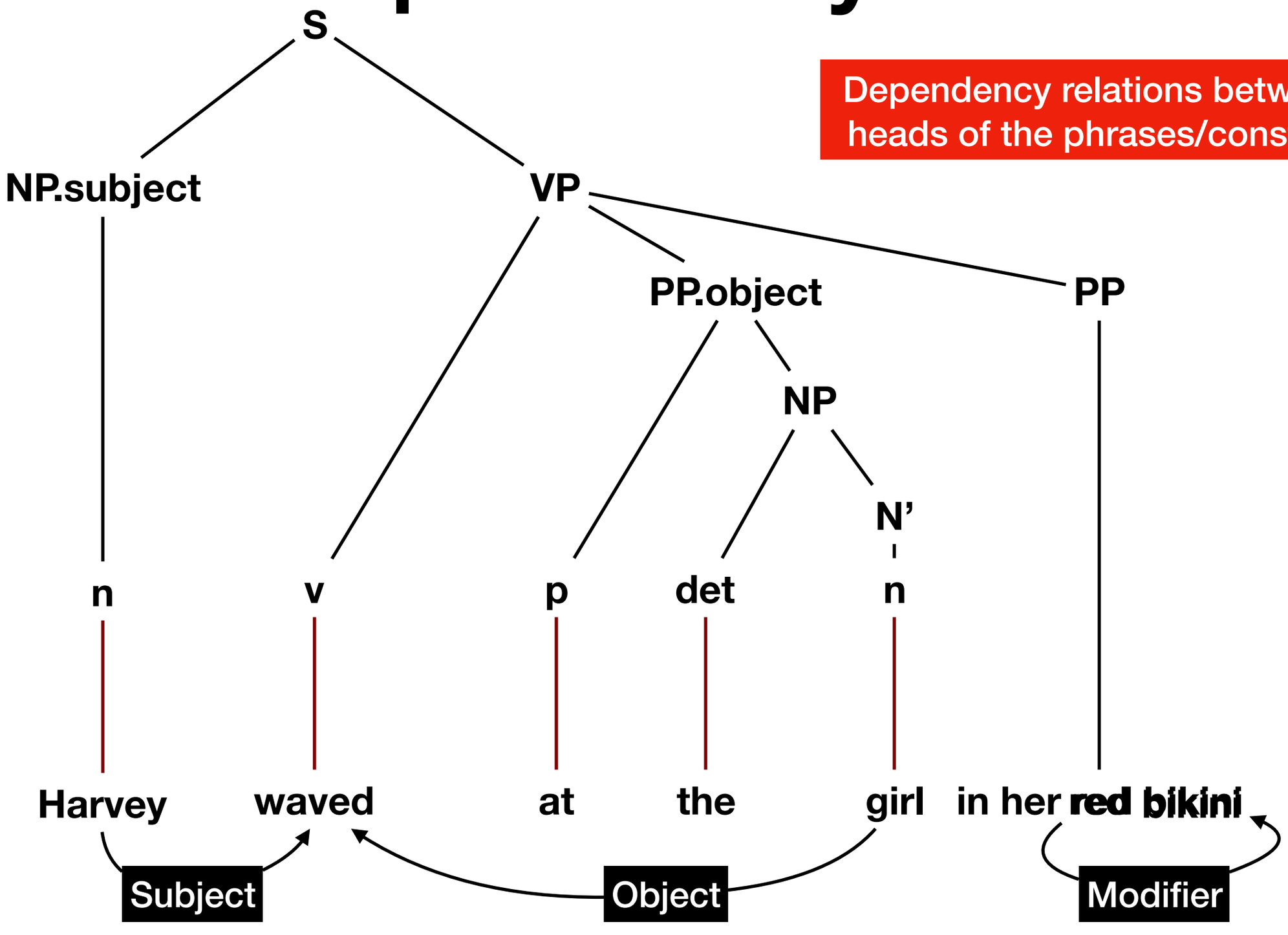
Syntactic trees



Syntactic functions

- Grammatical **Subject**: agreement with the main verb
    - the *boys* wave at the girl
    - the *books* were given to me
    - the *boys* were hit by the girl
    - the *girl* hits the boys
- Grammatical **Objects**: obligatory NPs or PPs to form a grammatical sentence
    - *the boys give the girl
    - *the boys fancy
    - *the boys treat the girl
- https://universaldependencies.org/u/dep/

Syntax Tree with dependency labels



Dependency relations between the heads of the phrases/constituents

Syntax

- Most important types of predicates in terms of obligatory arguments (the complementation=that what is needed to obtain a grammatical structure)

Valency	Predicate	Complementation			Example
Intransitive	walk.v	NP.subject			The cow walks
Transitive	perform.v	NP.subject	NP.direct object		The cow performs a looping
	count.v	NP.subject	PP(on).pp-object		The cow is hoping for a big applause
	be.v	NP.subject	NP.object | AP.object		This cow is a phenomenon. / This cow is phenomenal
Ditransitive	give.v	NP.subject	NP.direct object	NP.indirect object	The cow gives the spectators an unforgettable day

- A lexicon provides a list of verbs with their complementation patterns

Phrase structure parser

- Lookup words from a sentence in a sentence to find a candidate for a main verb
- Get the obligatory arguments of the verb
- Match the structure of surrounding phrases with the structure of the arguments (taken word order into account)
- Match the remaining phrases as non-obligatory elements
- If nothing left, a potential sentence structure is found

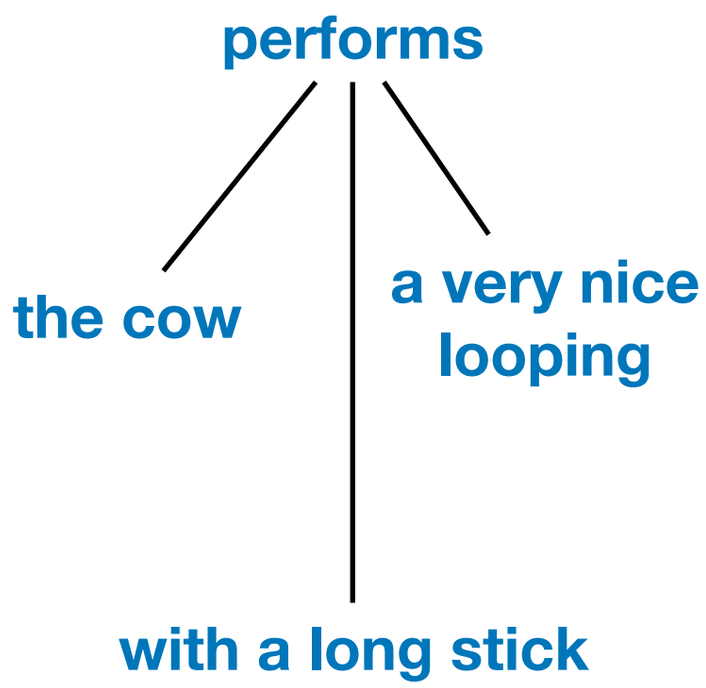
Syntax in short

- *The cow performs a very nice looping with a long stick.*

- The main verb is the centre of the sentence.

- The main verb gives you the obligatory arguments to make a grammatical sentence

- Next to the obligatory arguments there can also be optional adjuncts



More details

Syntax: some issues

- Constituents (S, VP, NP, PP, AP, AdvP) can be very small and infinitely large
    - He (NP); The nice green dogs with hats on their head (NP)
- PP-attachment ambiguity is often semantic or context dependent
    - Groucho shot an elephant in his pants.
- Scope is often semantic or context dependent
    - Old men and women can be annoying.
- Argument or adjunct depends on semantics or context
    - I count on your computer.
- A these can be easily combined in one sentence
    - Old men and women may be counting on their computers in their pants.

More details

Syntax: some issues

- Sentences are often ungrammatical!
    - My trainer don't tell me nothing between rounds. I don't allow him to. All I want to know is did I win the round.
        - Typo’s and somtimez not even a verb
- It's out of the date and lacks validity, so formats of late exams are quite different from its. But for foreign language learners, 'there-insertion' is quite handful.
- => A parser should be robust.

Parsers

- https://www.nltk.org/book/ch08.html
- Context free grammars use rewrite rules:

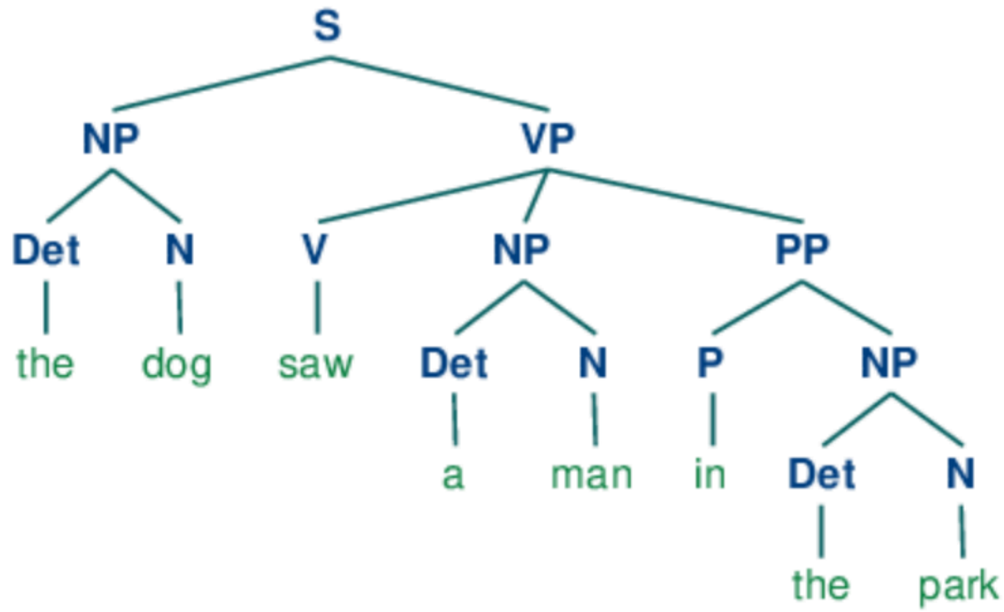
```
S -> NP VP
VP -> V NP | V NP PP
PP -> P NP
V -> "saw" | "ate" | "walked"
NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
Det -> "a" | "an" | "the" | "my"
N -> "man" | "dog" | "cat" | "telescope" | "park"
P -> "in" | "on" | "by" | "with"
```

Context Free Grammar

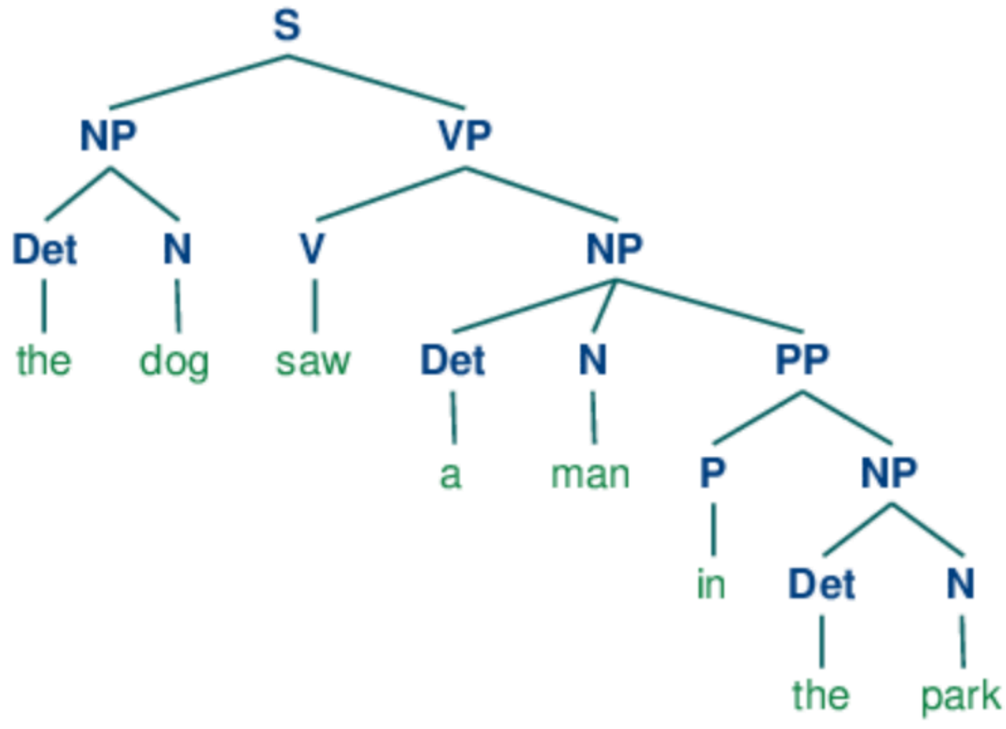
```
S -> NP VP
  VP -> V NP | V NP PP
  PP -> P NP
  V -> "saw" | "ate" | "walked"
  NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
  Det -> "a" | "an" | "the" | "my"
  N -> "man" | "dog" | "cat" | "telescope" | "park"
  P -> "in" | "on" | "by" | "with"
```

Symbol	Meaning	Example
S	sentence	*the man walked*
NP	noun phrase	*a dog*
VP	verb phrase	*saw a park*
PP	prepositional phrase	*with a telescope*
Det	determiner	*the*
N	noun	*dog*
V	verb	*walked*
P	preposition	*in*

(9) a.



b.



Stochastic parser

- Most parsers are trained from so-called treebanks: large collection of manually created parser trees

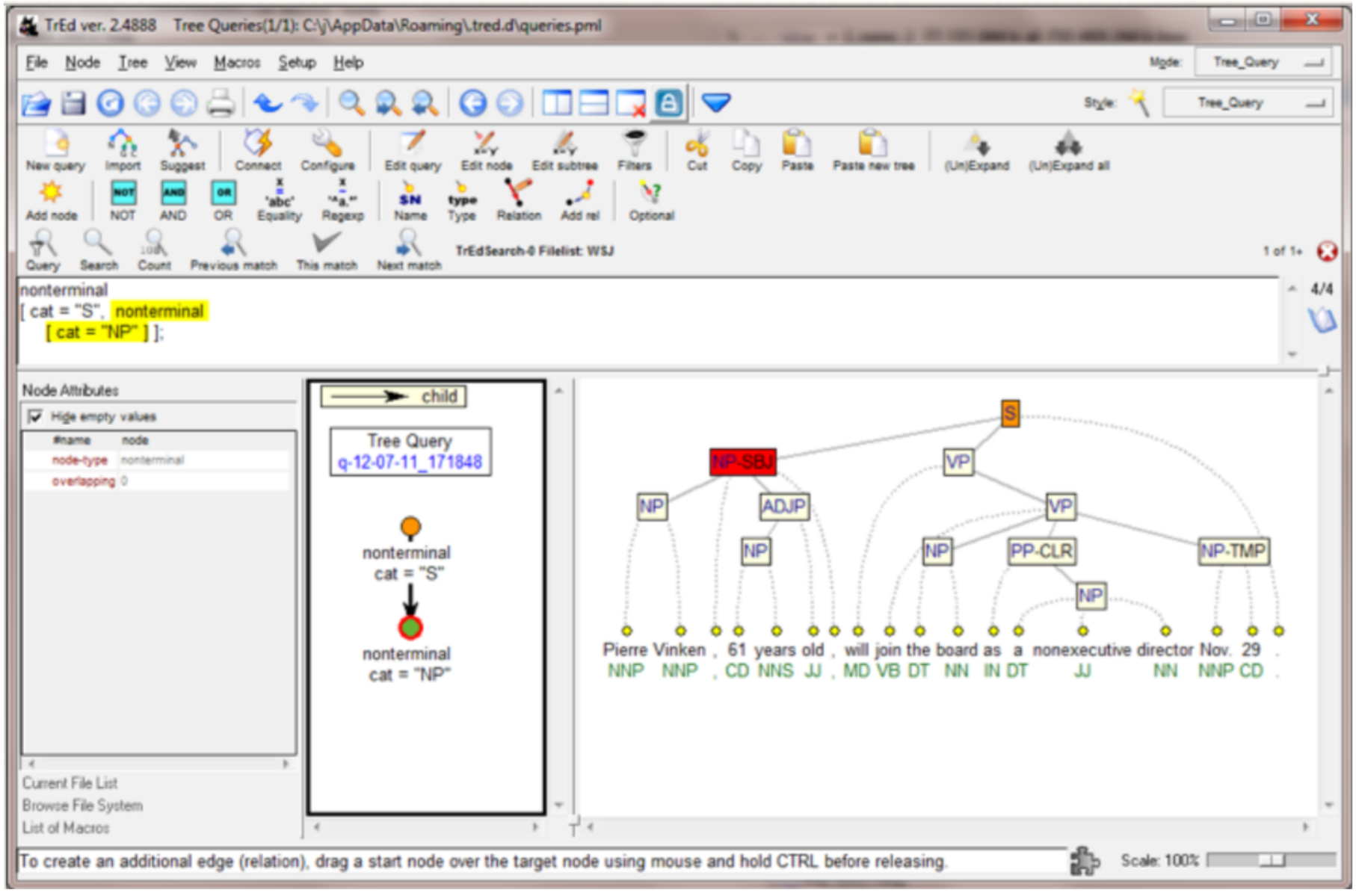
- Usually take tokenised text as input that is also lemmatised and tagged with parts-of-speech
    - Chunkers: only mark phrase structure boundaries
    - Phrase structure parsers add dependency relations
    - Dependency parser label dependencies

- Large quantities of annotated texts are needed

Treebanks

- Big set of parse trees, often created to train parsers on
  - Penn treebank: http://www.cis.upenn.edu/~treebank
  - Prague dependency treebank: http://ufal.mff.cuni.cz/pdt2.0
  - TiGer treebank: http://www.ims.uni-stuttgart.de/forschung/resourcen/korpora/tiger.en.html
- Penn Treebank:
  - originally just phrase structure, converted to dependencies by Collins (1996)
  - Currently contains basic predicate-argument structure

Penn treebank



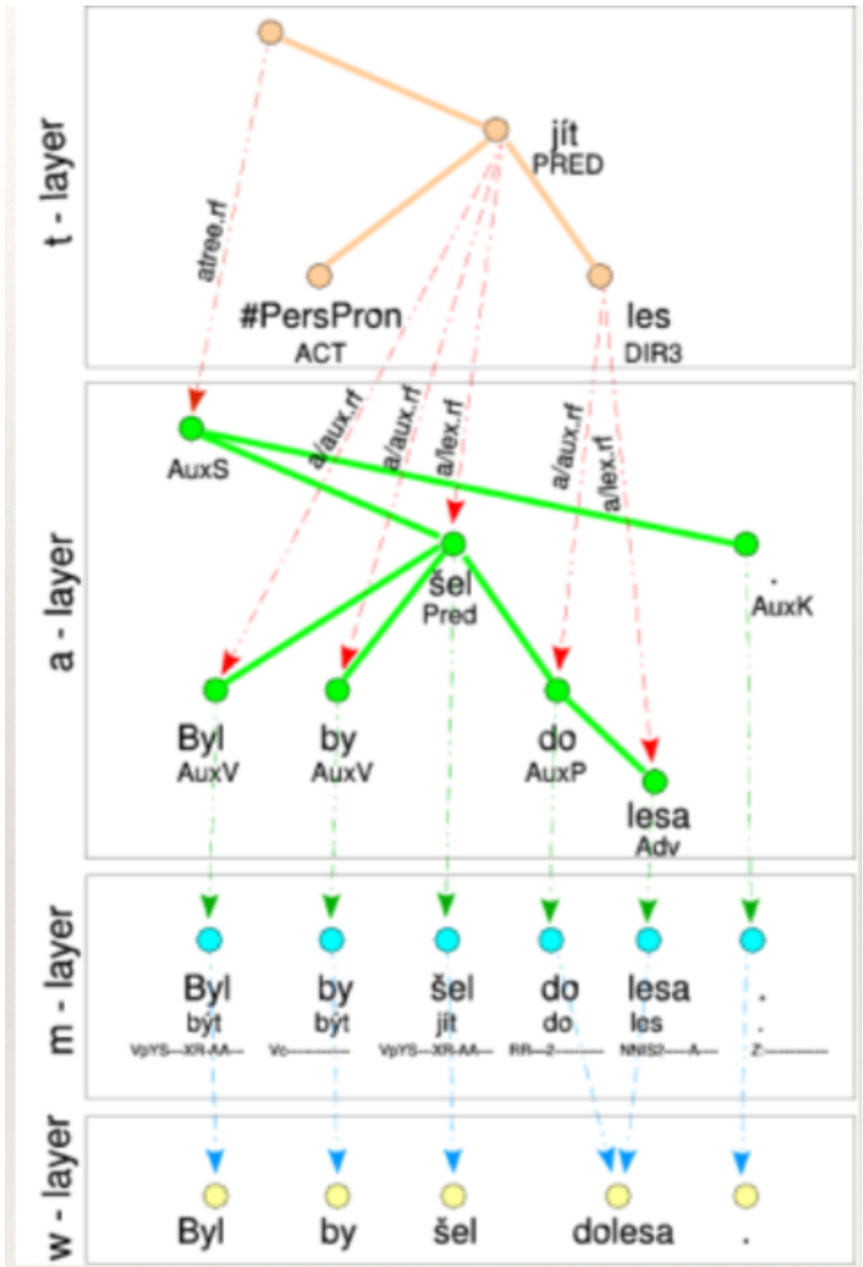
http://faculty.washington.edu/fxia/LAWVI/workshop_presentation_slides/special_session/pml/bak06-pmltq-q.PNG

Penn Treebank format

nltk_data/corpora/treebank/parsed/wsj_0003.prd

```
( (S (PP-TMP In
           (NP July))
     ,
     (NP-SBJ the Environmental Protection Agency)
     (VP imposed
        (NP a gradual ban)
      (PP-CLR on
           (NP (NP (ADJP virtually all)
                 uses)
               (PP of
                (NP asbestos)))))
.))
```

Prague Dependency Treebank



http://faculty.washington.edu/fxia/LAWVI/workshop_presentation_slides/special_session/pml/dolesa.png

Dependency Treebank format

nltk_data/corpora/dependency_treebank/wsj_0016.dp

```
1   The     DT 3
2   monthly   JJ 3
3   sales   NNS   4
4   have   VBP   0
5   been   VBN   4
6   setting     VBG   5
7   records     NNS   6
8   every   DT 9
9   monthNN     6
10 since   IN 9
11 MarchNNP   10
12 .    .    4
```

CoNLL annotations

NLTK: nltk_data/corpora/conll2000/train

- CoNLL = Computational Natural Language Learning competition
- Created training and test data for many NLP tasks for various languages.
- Word tokens are listed on a separate line for each document.
- Annotation are added in columns separate by TABs
- IOB annotation style:
  - I = inside
  - O = outside
  - B = beginning

- Chancellor NNP O
- of IN B-PP
- the DT B-NP
- Exchequer NNP I-NP
- Nigel NNP B-NP
- Lawson NNP I-NP
- 's POS B-NP
- restated VBN I-NP
- commitment NN I-NP
- to TO B-PP
- a DT B-NP
- firm NN I-NP
- monetary JJ I-NP
- policy NN I-NP
- has VBZ B-VP
- helped VBN I-VP
- to TO I-VP
- prevent VB I-VP
- a DT B-NP
- freefall NN I-NP
- in IN B-PP
- sterling NN B-NP
- over IN B-PP
- the DT B-NP
- past JJ I-NP
- week NN I-NP
- . . O

Stanford Parser

- Widely used statistical parser trained on the Penn Treebank
- PCFG: Probabilistic Context Free Grammar
- Also provides labels of dependencies
- Try it out: http://nlp.stanford.edu:8080/parser/index.jsp
- Performance:
    - F_1 = .85 for phrase structures
    - F_1 = .80 for dependency labelling

Stanford output

Tagging

```
Krelis/NNS  waved/VBD  at/IN  the/DT  girl/NN  with/IN  the/DT  bikini/NN  ./.
```

Parse

```
(ROOT
  (S
    (NP (NNS Krelis))
    (VP (VBD waved)
      (PP (IN at)
        (NP (DT the) (NN girl)))
      (PP (IN with)
        (NP (DT the) (NN bikini))))
    (. .)))
```

Stanford output

Typed dependencies

```
nsubj(waved-2, Krelis-1)
root(ROOT-0, waved-2)
prep(waved-2, at-3)
det(girl-5, the-4)
pobj(at-3, girl-5)
prep(waved-2, with-6)
det(bikini-8, the-7)
pobj(with-6, bikini-8)
```

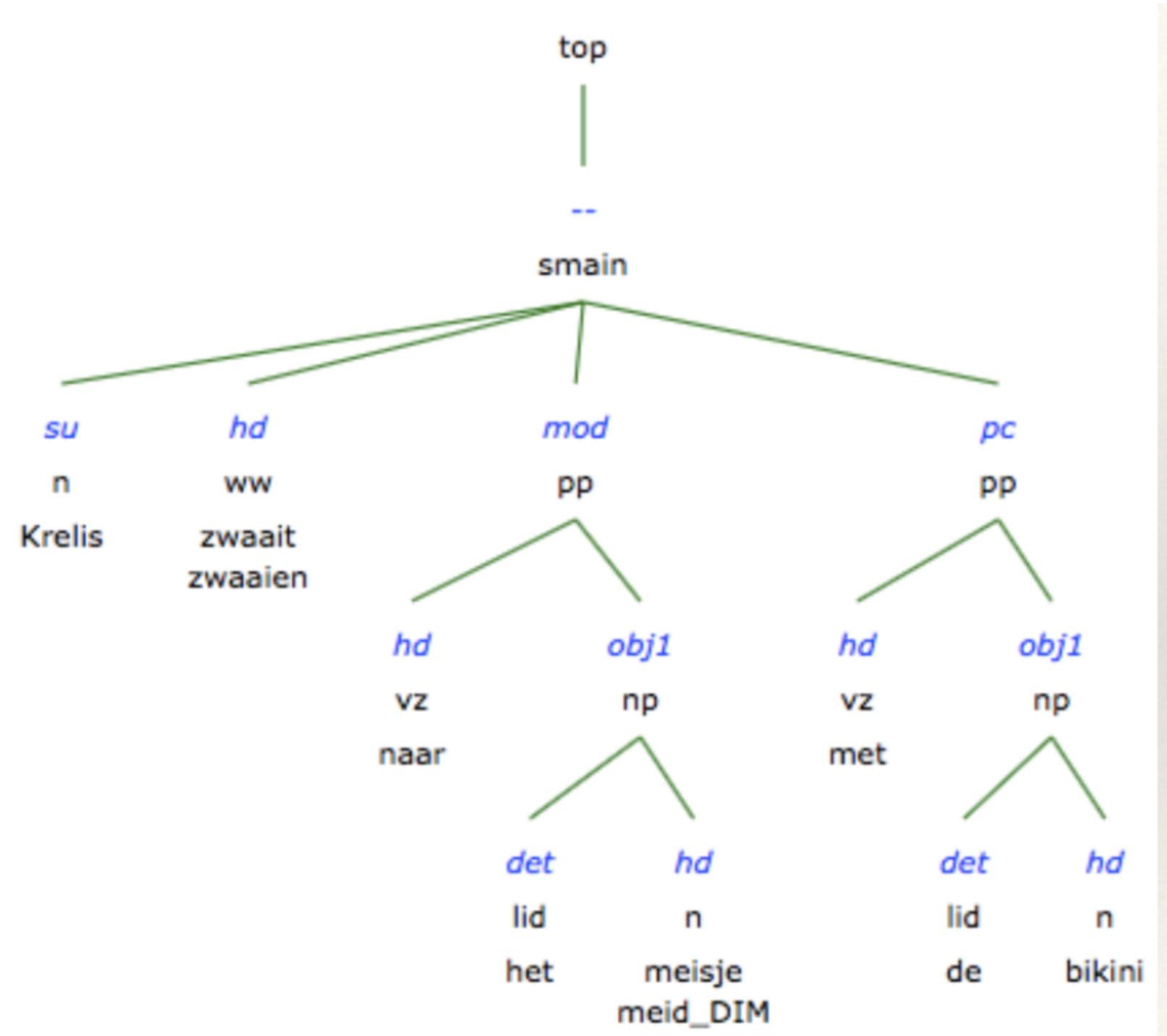
Typed dependencies, collapsed

```
nsubj(waved-2, Krelis-1)
root(ROOT-0, waved-2)
det(girl-5, the-4)
prep_at(waved-2, girl-5)
det(bikini-8, the-7)
prep_with(waved-2, bikini-8)
```

Alpino

- HPSG-based grammar for Dutch
- available at: http://www.let.rug.nl/vannoord/alp/Alpino
- Provides a set of all possible solutions
- Probabilistic parse ranking to find the most likely parse
- Rich output
- Accuracy: ~80%

Alpino example



Alpino format

nltk_data/corpora/alpino/alpino.xml

```
<alpino_ds version="1.2" id="0008">
  <node begin="0" cat="top" end="9" id="0" rel="top">
    <node begin="0" cat="inf" end="8" id="1" rel="--">
      <node begin="2" end="3" id="2" pos="verb" rel="hd" root="doe" word="doen"/>
      <node begin="0" cat="np" end="8" id="3" rel="obj1">
        <node begin="0" end="1" id="4" pos="noun" rel="hd" root="niks" word="Niks"/>
        <node begin="1" cat="ap" end="8" id="5" rel="mod">
          <node begin="1" end="2" id="6" pos="adj" rel="hd" root="anders" word="anders"/>
          <node begin="3" cat="cp" end="8" id="7" rel="obcomp">
            <node begin="3" end="4" id="8" pos="comparative" rel="cmp" root="dan" word="dan"/>
            <node begin="4" cat="inf" end="8" id="9" rel="body">
              <node begin="4" end="5" id="10" pos="adv" rel="mod" root="almaar" word="almaar"/>
              <node begin="5" cat="np" end="7" id="11" rel="obj1">
                <node begin="5" end="6" id="12" pos="adj" rel="mod" root="ruw" word="ruw"/>
                <node begin="6" end="7" id="13" pos="noun" rel="hd" root="materiaal" word="materiaal"/>
              </node>
              <node begin="7" end="8" id="14" pos="verb" rel="hd" root="verzamel" word="verzamelen"/>
            </node>
          </node>
        </node>
      </node>
    </node>
    <node begin="8" end="9" id="15" pos="punct" rel="--" root="." word="."/>
  </node>
  <sentence>Niks anders doen dan almaar ruw materiaal verzamelen .</sentence>
</alpino_ds>
```

Parsing: good to know

- Results for Stanford & Alpino apply to clean newspaper text

- Parsing is expensive in memory and time

- Challenges:

    - linguistic phenomena such as conjunctions, ellipsis & long distance dependencies

    - problems in tokenisation and PoS-tagging can harm the parser

If you don’t need full parse trees: Chunking

- Chunking (also called “shallow parsing”) provides a cheap and robust alternative to parsing
- Chunks are constituents
- Chunkers do not provide full syntax trees, usually only constituents up to a certain level in depth (typically 2)
- After chunking a classifier can assign phrase types as well
- [Krelis]NP zwaaide [naar [het meisje]NP]PP [met [de bikini]NP]PP

Words have meanings

- https://en.wikipedia.org/wiki/Head_(disambiguation)
- http://wordnetweb.princeton.edu/perl/webwn?c=7&sub=Change&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&i=-1&h=0000000000000000000000000000000000&s=head
- https://nl.wikipedia.org/wiki/Hoofd
- https://www.vandale.nl/gratis-woordenboek/nederlands/betekenis/hoofd#.XPDv5aeB0Wo
- http://cornetto.clarin.inl.nl/simple_search.xql?word=hoofd&pos=&p_length=20&offset=0&search_type=simple

Head (disambiguation)

From Wikipedia, the free encyclopedia

The **head** (**Human head**) is the part of an animal or human that usually includes the brain, eyes, ears,

Head may also refer to:

Arts, entertainment, and media [edit]

Music [edit]

Albums [edit]

- *Heads* (Bob James album), 1977
- *Head* (The Jesus Lizard album), 1990
- *Head* (the Monkees album), a 1968 soundtrack of the movie
- *Heads* (Osibisa album), 1972

Songs [edit]

- *Head* (The Cooper Temple Clause song), track from *Make This Your Own*
- "Head" (Julian Cope song), 1991
- "Head" (Prince song)
- "Head", a song by Mark Lanegan from *Bubblegum*
- "Head", a song by Static-X from *Beneath... Between... Beyond...*
- "Head", a song by Todd Sheaffer from *The Black Bear Sessions* and *Elko*
- "Head", a song by Lotion from *full Isaac*
- "Heads", a song by Hawkwind from *The Xenon Codex*

Other music [edit]

- Head (band), an English rock band
- The Head (band), an indie rock band from Atlanta, Georgia
- Head (music), a main theme in jazz
- Drumhead, a membrane on a drum
- Headstock, a part of an instrument

Film and television [edit]

- *Head* (film), a 1968 film starring The Monkees
- *Heads* (film), a 1994 TV movie
- *The Head* (film), a 1959 German horror film directed by Victor Trivas
- *The Head*, a 1994–1996 American animated television series
- "Head" (*Blackadder*), a 1986 episode of *Blackadder*
- Head (*American Horror Story*), a 2013 episode of the anthology television series

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: head 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"
Display options for word: word#sense number

Noun

- S: (n) **head#1**, caput#2 (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head#2** (a single domestic animal) *"200 head of cattle"*
- S: (n) mind#1, **head#3**, brain#3, psyche#1, nous#2 (that which is responsible for one's thoughts, feelings, and conscious brain functions; the seat of the faculty of reason) *"his mind wandered"; "I couldn't get his words out of my head"*
- S: (n) **head#4**, chief#1, top dog#1 (a person who is in charge) *"the head of the whole operation"*
- S: (n) **head#5** (the front of a military formation or procession) *"the head of the column advanced boldly"; "they were at the head of the attack"*
- S: (n) **head#6** (the pressure exerted by a fluid) *"a head of steam"*
- S: (n) **head#7** (the top of something) *"the head of the stairs"; "the head of the page"; "the head of the list"*
- S: (n) fountainhead#2, headspring#1, **head#8** (the source of water from which a stream arises) *"they tracked him back toward the head of the stream"*
- S: (n) **head#9**, head word#2 ((grammar) the word in a grammatical constituent that plays the same grammatical role as the whole constituent)
- S: (n) **head#10** (the tip of an abscess (where the pus accumulates))
- S: (n) **head#11** (the length or height based on the size of a human or animal head) *"he is two heads taller than his little sister"; "his horse won by a head"*
- S: (n) capitulum#1, **head#12** (a dense cluster of flowers or foliage) *"a head of cauliflower"; "a head of lettuce"*
- S: (n) principal#2, school principal#1, head teacher#1, **head#13** (the educator who has executive authority for a school) *"she sent unruly pupils to see the principal"*
- S: (n) **head#14** (an individual person) *"tickets are $5 per head"*
- S: (n) **head#15** (a user of (usually soft) drugs) *"the office was full of secret heads"*
- S: (n) promontory#1, headland#1, **head#16**, foreland#1 (a natural elevation (especially a rocky one that juts out into the sea))
- S: (n) **head#17** (a rounded compact mass) *"the head of a comet"*
- S: (n) **head#18** (the foam or froth that accumulates at the top when you pour an effervescent liquid into a container) *"the beer had a large head of foam"*
- S: (n) forefront#1, **head#19** (the part in the front or nearest the viewer) *"he was in the forefront"; "he was at the head of the column"*
- S: (n) pass#9, **head#20**, straits#2 (a difficult juncture) *"a pretty pass"; "matters came to a head yesterday"*
- S: (n) headway#2, **head#21** (forward movement) *"the ship made little headway against the gale"*
- S: (n) point#20, **head#22** (a V-shaped mark at one end of an arrow pointer) *"the point of the arrow was due north"*
- S: (n) question#2, **head#23** (the subject matter at issue) *"the question of disease merits serious discussion"; "under the head of minor Roman poets"*
- S: (n) heading#1, header#1, **head#24** (a line of text serving to indicate what the passage below it is about) *"the heading seemed to have little to do with*

Hoofd

Hoofd kan verwijzen naar:

- hoofd (anatomie) (bij dieren ook *kop*), het bovenste deel van het lichaam
- het hoogste of voorste deel (vgl. Aan het hoofd staan.)
- manager of leidinggevende
- hoofdje, een bloeiwijze
- (per) hoofd van de bevolking, zie per capita
- Krib (rivier), synoniem voor een korte stenen dam in een rivier
- sluishoofd
- hoofd (taalkunde), een term uit de ontleding
- Hoofd (Hoorn), een straat in Hoorn

Betekenis 'hoofd'

Je hebt gezocht op het woord: hoofd.

hoofd *(het; o;* meervoud: *hoofden)*

1 bovenste deel van het menselijk lichaam: *aan iets het hoofd bieden* zich ertegen verzetten; *iemands hoofd eisen* zijn aftreden eisen; *een hard hoofd in iets hebben* een zaak somber inzien; *heel wat aan zijn hoofd hebben* de zorg voor veel dingen hebben; *er hangt ons iets boven het hoofd* er dreigt gevaar; *iem., iets over het hoofd zien* (per ongeluk) niet zien; *uit het hoofd leren* van buiten; *iem. voor het hoofd stoten* kwetsend behandelen; *zich het hoofd breken over iets* erover tobben; *het groeit me boven het hoofd* ik kan het niet meer overzien, het wordt me te veel; *het hoofd koel houden* rustig blijven, niet in paniek raken; *zijn hoofd stoten (a)* het door stoten bezeren; *(b)* afgaan, gezichtsverlies lijden; *het hoofd boven water houden (a)* niet onder de omstandigheden bezwijken; *(b)* zich financieel redden; *uit hoofde van* wegens

2 verstand: *niet goed bij het hoofd zijn* min of meer gek

3 eerste, met leiding belaste, voornaamste persoon

4 persoon: *zoveel hoofden, zoveel zinnen* zoveel mensen, zoveel zienswijzen; *zestig euro per hoofd*

5 het bovenste, voorste gedeelte van iets: *het hoofd van een brief*; *aan het hoofd staan* de leiding hebben

Cornetto Demo

Simple Search | Advanced Search | Synset Visualization | Help

Simple Search | Matching Lexical Entries

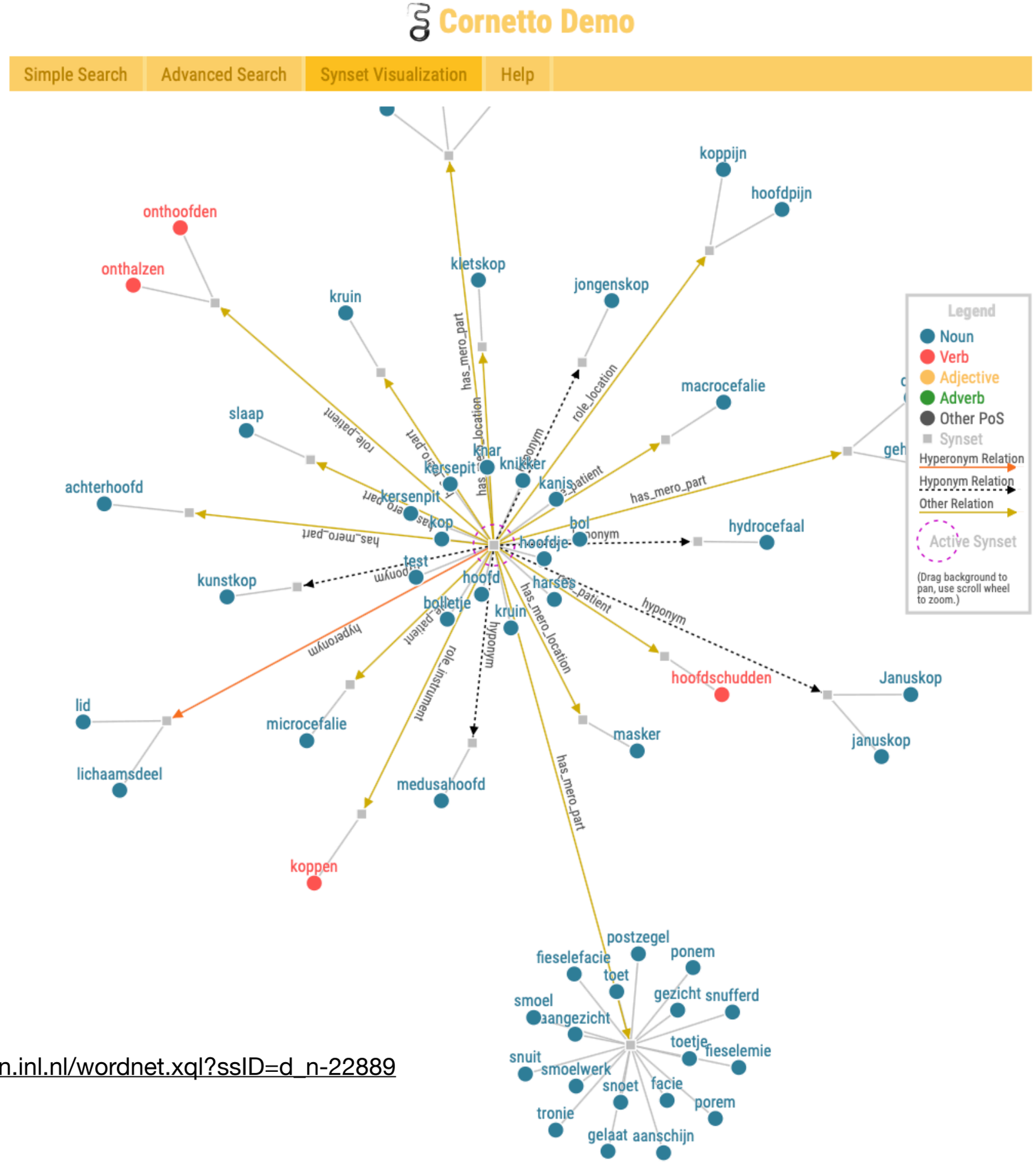
Matching Lexical Entries for 'hoofd'

Download Matches as LMF-XML (Max. 500)

1 to 20 from 46 Previous Next

ID	Written Form	Semantics	PoS	Examples	Pragmatics Syntax
hoofd-n-1	hoofd	lichaamsdeel dat op de nek staat	noun	een gat in je hoofd vallen (4 more)	anatomy
hoofd-n-2	hoofd	leider	noun	het hoofd van de school (3 more)	
hoofd-n-3	hoofd	hersens als denkvermogen	noun	je moet niet alleen met je hoofd werken, maar ook met je hart (1 more)	
hoofd-n-4	hoofd	voorste/bovenste deel	noun	(ze zat) aan het hoofd van de tafel	
hoofd-mwe-51628	iets uit het hoofd leren/kennen	iets van buiten leren/kennen	-		
hoofd-mwe-51627	hoe haal je het in je hoofd?	hoe kom je er in hemelsnaam bij om zoiets te doen?	-		
hoofd-mwe-51626	mijn hoofd loopt om	ik heb het ontzettend druk	-		
hoofd-mwe-51625	het hoofd vol hebben van iets	steeds aan iets denken	-		
hoofd-mwe-r_n-17064-31	het hoofd laten hangen	de moed opgeven	-		
hoofd-mwe-r_n-17064-30	iemand/iets het hoofd bieden	je verzetten tegen iemand/iets	-		
hoofd-mwe-r_n-17064-29	iemand iets naar het hoofd slingeren	iemand iets verwijten in een ruzie	-		
hoofd-mwe-r_n-17064-28	iemand voor het hoofd stoten	iemand beledigen	-		
hoofd-mwe-51629	zich het hoofd breken over iets	door diep nadenken proberen een oplossing voor iets te vinden	-		
hoofd-mwe-51630	mijn hoofd staat er niet naar	ik ben er niet voor in de stemming	-		
hoofd-mwe-51631	iets uit je hoofd	iets niet doen omdat het	-	als je dat maar uit je hoofd laat!	

Cornetto Synsets



- http://cornetto.clarin.inl.nl/wordnet.xql?ssID=d_n-22889

Lexical structures
http://wordpress.let.vupr.nl/cornetto/

```
<LexicalEntry id="hoofd-n-2" partOfSpeech="noun">
  <Lemma writtenForm="hoofd"/>
  <WordForms>
    <WordForm writtenForm="hoofd" grammaticalNumber="singular" article="het"/>
    <WordForm writtenForm="hoofden" grammaticalNumber="plural" article="de"/>
  </WordForms>
  <Morphology/>
  <MorphoSyntax pronominalAndGrammaticalGender="m_f"/>
  <Sense senseId="r_n-17065" definition="leider" synset="d_n-34121" origin="">
    <SenseRelations><SenseGroup relationType="co-annotation" targetSenseId=“r_n-17066"/></SenseRelations>
    <Semantics-noun reference="common" countability="count" semanticType="human"/>
    <Pragmatics/>
    <SenseExamples>
      <SenseExample id="53582">
        <canonicalForm canonicalform="het hoofd van de school" phraseType="np" expressionType="freeCombination"/>
        <Syntax_ex>
          <combiWord lemma="school" partOfSpeech="noun"/>
        </Syntax_ex>
        <Semantics_ex/>
        <Pragmatics/>
      </SenseExample>
      <SenseExample id="53583">
        <canonicalForm canonicalform="(de paus is) het hoofd van de rooms-katholieke kerk"
                       phraseType="sentence" expressionType="freeCombination"/>
        <Syntax_ex>
          <combiWord lemma="kerk" partOfSpeech="noun"/>
        </Syntax_ex>
        <Semantics_ex/>
        <Pragmatics/>
      </SenseExample>
    </SenseExamples>
  </Sense>
</LexicalEntry>
```

Word meaning

- **polysemy**
    - related meanings: metonymic, metaphoric, specialisation
    - school, horse, to support, strong, chicken
- **underspecification or generalisation**
    - person (male? female? adult? child?)
    - animal
- **homonymy**
    - unrelated meanings, arbitrary form overlap
    - tear (verb/noun), bark (dog, ship),
    - bank (verb/noun/furniture/finance)

More details

Variation

How many words for a thing?

5000 words for person

More details

mortal; posturer; controller; **withholder**; suppressor; subduer; fugitive; divider; subdivider; outcaste; bereaved_person; yielder; nude; streaker; **unperson**; baby_buster; neighbor; loose_cannon; ladino; communicator; announcer; town_crier; caller; caller; muezzin; hisser; gossipmonger; yenta; cat; telltale; **scandalmonger**; allegoriser; presenter; promisee; quoter; transmitter; spammer; answerer; hedger; assenter; interviewee; **don’tknow**; testee; passer; popularizer; avower; laudator; clapper; waffler; wirer; conferrer; confessor; reporter; newswoman; television_reporter; anchorman; avower; postulator; author; gagwriter; poet; sonneteer; poetess; homer; elegist; frost; key; gilbert; gray; pound; poet_laureate; odist; spender; poet_laureate; bard; biographer; autobiographer; hagiographer; novelist; folk_writer; folk_poet; **cyberpunk**; west; wood; authoress; abstractor; pamphleteer; speechwriter; drafter; paragrapher; space_writer; tragedian; snow; day; playwright; rice; kid; cooper; buck; scriptwriter; film_writer; wordmonger; framer; literary_hack; rand; coauthor; scenarist; litterateur; **word-painter**; wordsmith; alliterator; sand; e._e._cummings; heller; grass; spark; jour nalist; photojournalist; reed; stone; sob_sister; sports_writer; newspaperwoman; war_correspondent; foreign_correspondent; broadcast_journalist; gazetteer; columnist; newspaper_columnist; newspaper_critic; gossip_columnist; agony_aunt; scribbler; rhymer; lyrist; rice; ghostwriter; librettist; compiler; encyclopaedist; lexicologist; etymologist; synonymist; neologist; polemist; commentator; contributor; twaddler; alarmist; stirrer; letter_writer; pen_pal; broadcaster; telecaster; announcer; sportscaster; tv_announcer; newscaster; news_reader; radio_announcer;

More details

- naarling, beroerling, ellendeling, etterbak, etterbuil, fielt, fluim, gemenerik, hond, hondenlul, kankerlijer, kelerelijder, kelerelijer, klerelijer, kloot, kloothommel, klootspiraal, klootzak, kwal, lamgat, lammeling, lamstraal, lamzak, lazersteen, lazerstraal, loeder, lul, lulhannes, lulletje, miesgasser, mispunt, onverlaat, paardelul, paardenlul, patjakker, pleurislijder, ploert, plurk, pokkenlijer, pokkenvent, pooier, rasploert, reptiel, rotzak, schoelje, schoft, serpent, smeerlap, stinker, teringlijder, tyfuslijer, vuilak, zakkenwasser, zwijn, zak, hondelul, etter, lelijkerd, smiecht, pokkenlijder, sekreet, stinkerd, individu
- huichelaar, Januskop, draaikont, farizeeër, hypocriet, januskop, jezuïet, smoelentrekker, valsaard, valserik, veinzaard, veinzer
- onruststoker, aanstoker, aanzetter, agitator, herrieschopper, onrustzaaier, oproerkraaier, opruier, paniekzaaier, provocateur, raddraaier, roervink, stemmingmaker, stokebrand, stoker, woelgeest
- boef, booswicht, galgeaas, galgebrok, galgenaas, gannef, kwaaddoener, satan, slechterik, snoodaard, spitsboef, zwijnjak, schurk
- krankzinnige, fanatiekeling, geesteszieke, gek
- dwaas, achterlijke, gek, halvezool, idioot, imbe
  mafketel, mafkikker, malloot, nar piechem, zot

8.643 words for a person

4,000 words for move

More details

go; swim; buoy; drive; island_hop; whistle; ski; slalom; hot-dog; wedel; water_ski; schuss; pass_over; breeze; err; return; revisit; retrace; cut_back; resurrect; return; home; head_home; double_back; bounce; boomerang; fly; come; retrograde; walk; constitutionalize; speed; bang; **swash**; tread; step_on; beetle; circulate; drift; swim; bucket_along; shoot; rip; barge; dash; plunge; hurtle; sit; canter; override; prance; ride_herd; ride_horseback; prance; post; trot; gallop; canter; gallop; outride; lance; scramble; plough; sift; **zigzag**; billow; pursue; haunt; tail; tree; hound; ferret; run_down; quest; stalk; roll; troll; bowl; travel_purposefully; wend; whisk; cruise; stooge; steamer; go_forward; head; make; trace; limp; wander; roar; ease; circulate; float; ride; shack; draw; caravan; career; raft; swap; thrash; retreat; cocoon; automobile; step; backpedal; blow; stream; tide; waft; crawl; formicate; slice_into; run; precede; lead; draw_away; travel_along; ascend; heel; turn; angle; push; travel; ride; fly; cruise; ship; sail; wind; snake; pan; repair; taxi; precess; cast; **jazz_around**; maunder; travel; itinerate; go_up; rise_up; resurface; well; intumesce; emerge; uprise; ferry; transfer; betake_oneself; march_on; string; edge; forge; penetrate; rachet_up; sneak_up; plough_on; draw_in; slide_by; fell; impinge; overtake; clear; hop; get_by; tram; prance; derail; go_through; get_across; stride; take; ford; tramp; jaywalk; crisscross; bridge; walk; hop; course; cut; muscle; lock; negociate; pass_through; make; jostle; bushwhack; claw; pass_over; cut; crash; transit; blunder; cycle; cycle_on; squeak_by; break_through; run; overstep; go_around; drive; pull; cut_in; wing; fly_on; soar; rack; buzz; hover; poise; flight; go_down; go_down; drip; correct; subside; dismount; pitch; go_down; founder; subside; submerge; dive; belly-flop; jackknife; flop; rope_down; cascade; drop; flump_down; decline; dip;

2,037 words for noise

More details

grinding; racket; report; squeak; clap; clack; snore; chatter; chattering; brouhaha; hubbub; uproar; katzenjammer; clatter; shrieking; scream; screech; screaming; shriek; screeching; blaring; blare; din; cacophony; clamor; grumble; rumble; grumbling; rumbling; squawk; plump; crackling; crackle; crepitation; decrepitation; snap; explosion; squish; rhonchus; hum; humming; swoosh; whoosh; clangoring; clang; clank; clash; clangor; crash; clangour; whisper; whispering; rustling; rustle; chug; sizzle; plonk; howl; squeal; plop; scrape; scratching; scratch; scraping; chatter; chattering; ding-dong

More details

Ambiguity is pervasive and we do not perceive it!!!

- 121 most frequent English nouns have on average 7.8 meanings each and account for about 20% of word occurrences in real text (in the Princeton WordNet (Miller 1990), according to Ng and Lee (1996)).

- The ambiguity demo: http://130.37.53.15:5001/

- “He gave a soft ball across the line from the center of the field, making a major point and giving a minor lead.”

More details

The meaning puzzle

Ambiguity demo

Sentence: He(2) gave(44) a soft(19) ball(12) across the line(30) from the center(18) of the field(17) , making(49) a maj

Compute

He(2) gave(44) a soft(19) ball(12) across the line(30) from the center(18) of the field(17) , making(49) a major(8) point(26) and giving(44) a minor(10) lead(17) . = 1404174924472320 0 possible meaning combinations

14,041,749,244,723,200

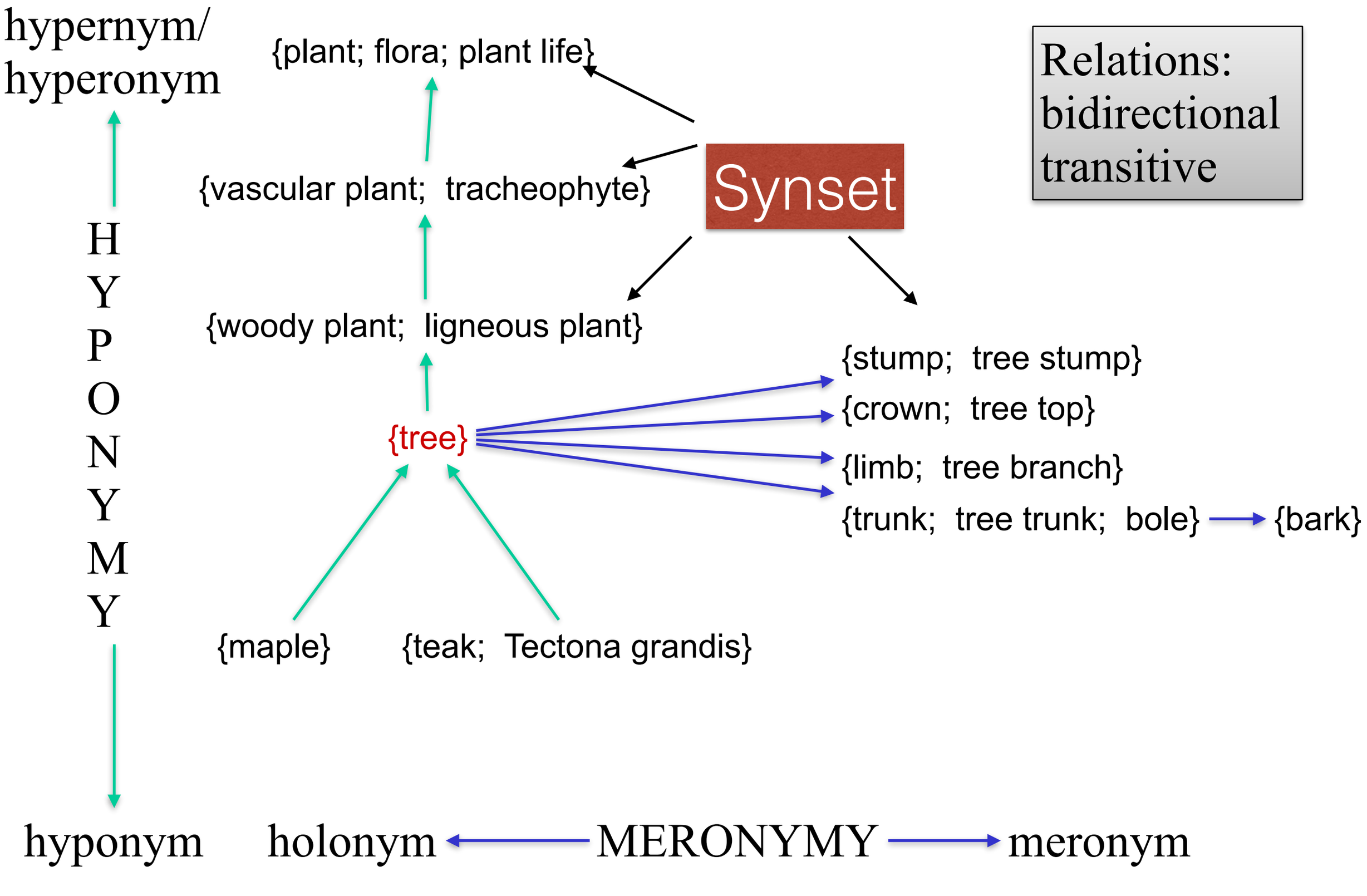
More details

Princeton WordNet

- Semantic lexical database based on psycholinguistic principles,Miller et al. 1991.
- Distinguishes words from concepts represented as synsets and glosses
- Words and synsets have relations
- Captures polysemy (ambiguity) and synonymy (variation)

WordNet

More details



More details

Wordnet 3.0 statistics

POS	Unique	Synsets	Total
	Strings		Word-Sense Pairs
Noun	117,798	82,115	146,312
Verb	11,529	13,767	25,047
Adjective	21,479	18,156	30,002
Adverb	4,481	3,621	5,580
Totals	155,287	117,659	206,941

Synonymy: two words shared a concept <plank, board>
Homonymy: one word belongs to two unrelated concepts <cell>
Polysemy: one word has more than one related concept <school>

More details

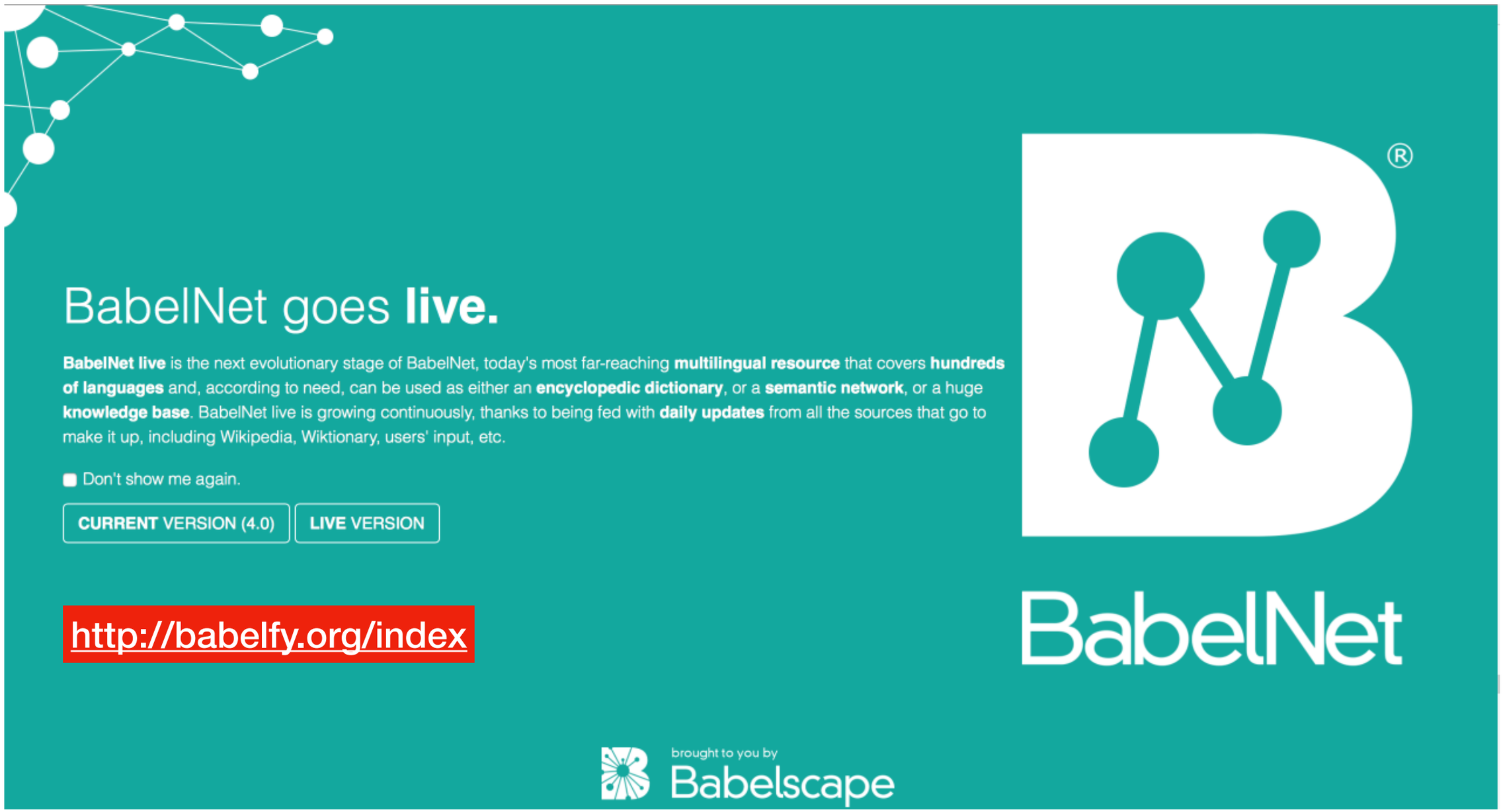
Wordnet 3.0 statistics

POS	Average Polysemy	Average Polysemy
	Including Monosemous Words	Excluding Monosemous Words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.4	2.71
Adverb	1.25	2.5

BabelNet: Combining WordNet & Wikipedia

More details

http://babelnet.org

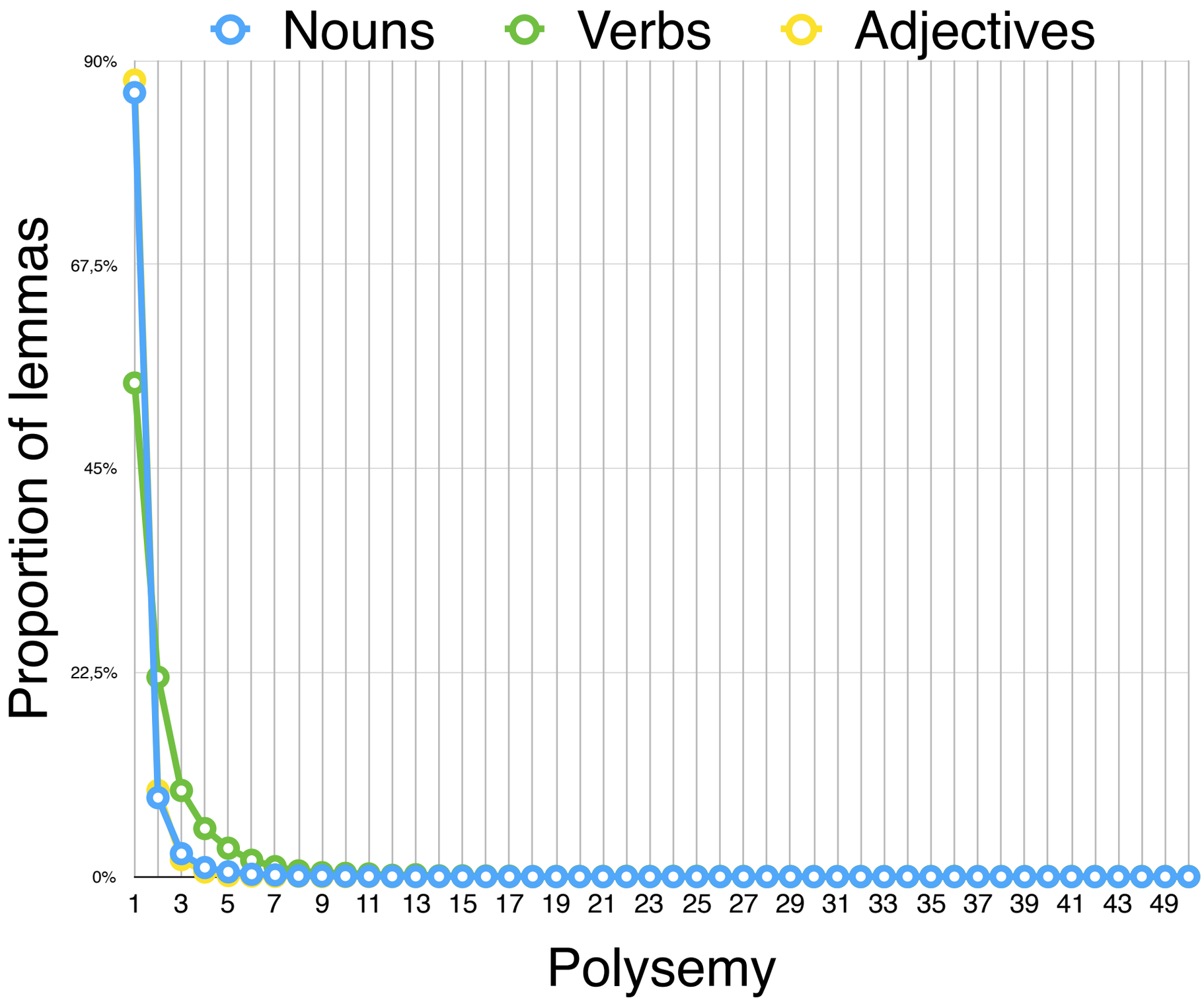


Give me a break!!!!

Top most polysemous words in WordNet3.0

Nouns	Polysemy	Verbs	Polysemy	Adjecti	More details
head	33	break	59	active	11
line	30	make	49	soft	8
point	26	give	44	light	8
base	20	take	42	inactive	8
case	20	cut	41	major	7
cut	20	run	41	minor	7
center	18	carry	40	open	6
field	17	draw	36	dry	6
lead	17	get	36	direct	6
shot	17	hold	36	hard	6
stock	17	play	35	heavy	6
play	17	fall	32	short	6
position	16	go	30	long	5
break	16	catch	29	critical	5
run	16	call	28	cut	5
place	16	work	27	straight	5
pass	16	raise	27	immature	5
form	16	turn	26	uncut	5
Service	15	cover	26	offensive	5
mark	15	check	25	national	5
order	15	pass	25	right	5
light	15	charge	25	syllabic	5
bar	15	set	25	single	4

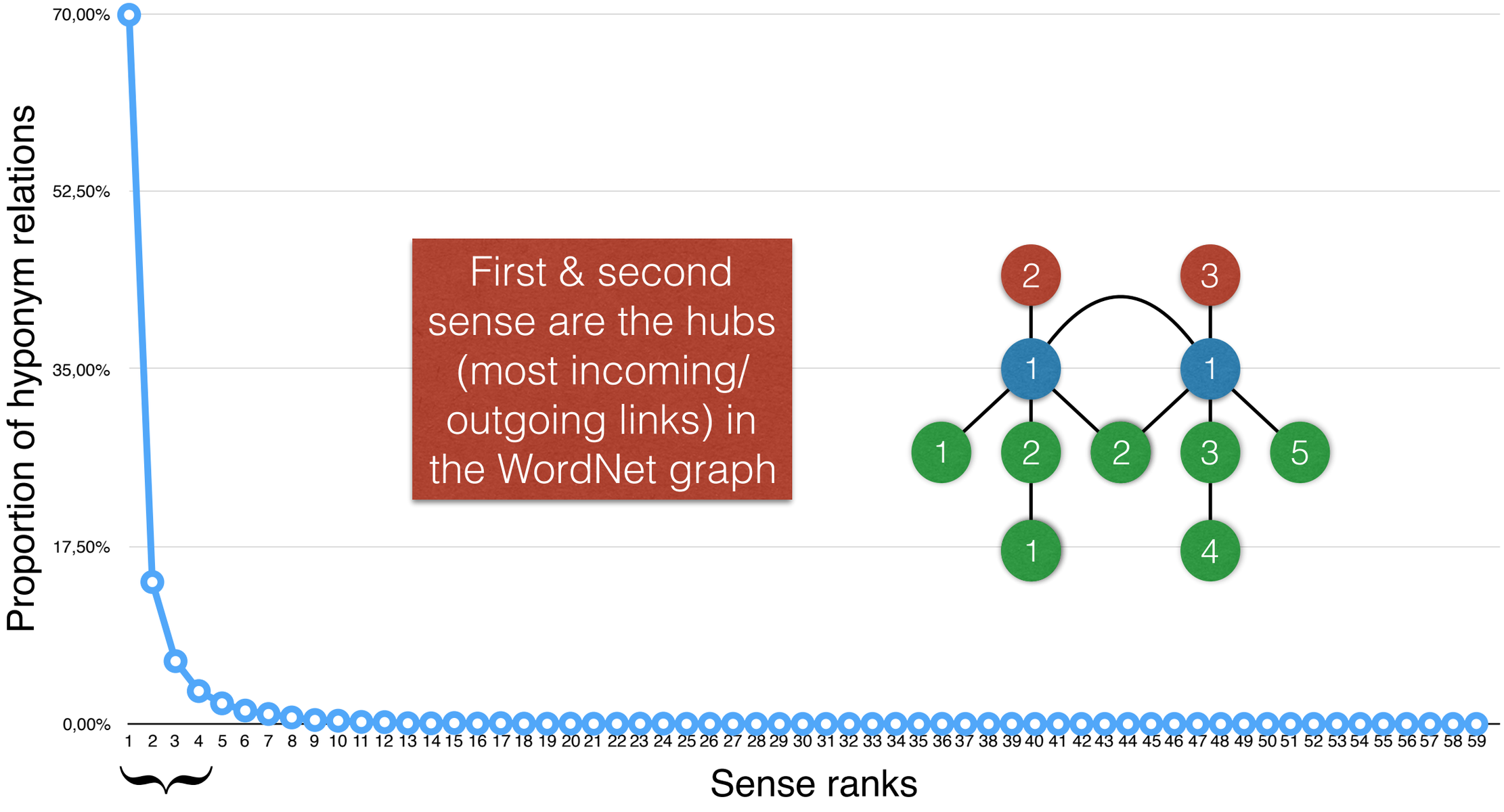
Polysemy in WordNet3.0



Pol.	Nouns	Verbs	Adj.
1	101863	6277	7313
2	10257	2536	791
3	2989	1095	156
4	1178	611	42
5	620	361	10
6	306	208	6
7	212	127	2
8	94	73	3
9	96	50	
10	60	41	
11	48	34	1
12	25	17	
13	14	24	
14	8	10	
15	10	10	
16	6	8	
17	5	6	
18	1	1	
59		1	
	117792	11490	8324

More details

Distribution of hyponyms over sense ranks in WordNet1.3



First & second sense are the hubs (most incoming/ outgoing links) in the WordNet graph

More details

From EuroWordNet to Global WordNet

- Currently, wordnets exist for more than 100 languages, including: Arabic, Bantu, Basque, Chinese, Bulgarian, Estonian, Hebrew, Icelandic, Japanese, Kannada, Korean, Latvian, Nepali, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, Zulu...
- Many languages are genetically and typologically unrelated
- http://www.globalwordnet.org

More details

Open Multilingual Wordnet

- Open source wordnets (200+)
- Expansion of the Princeton WordNet
- http://compling.hss.ntu.edu.sg/omw/

Sentences have meanings

- The father hit his sons with a hammer
  - the father = NP, subject, agent
  - his sons = NP, direct object, patient
  - with a hammer = PP, adjunct, instrument
- The sons were hit by his father with a hammer
  - ????
- grammatical constituent + syntactic function + semantic role

- **Agent:** performs with control (can stop doing it)
- **Patient:** undergoes the action and is changed by it
- **Instrument:** what the agent uses to perform the action
- **Others:** recipient, thema, source, path, goal …

From syntax to semantics

The boy ran from the shop across the street to his mummy

The boy fell

Harvey bought her flowers

She got flowers from Harvey

Flowers were given to her by Harvey

She broke Harveys eye socket

The hammer broke his eyesocket

His eye socket broke

What colors correspond with what semantic relations?

Agent, patient, theme, beneficiary/recipient, instrument,
Source, path, goal,

From syntax to semantics

The boy ran from the shop across the street to his mummy

The boy fell

Harvey bought her flowers

She got flowers from Harvey

Flowers were given to her by Harvey

She broke Harveys eye socket

The hammer broke his eyesocket

His eye socket broke

What colors correspond with what semantic relations?

Agent, patient, theme, beneficiary/recipient, instrument,

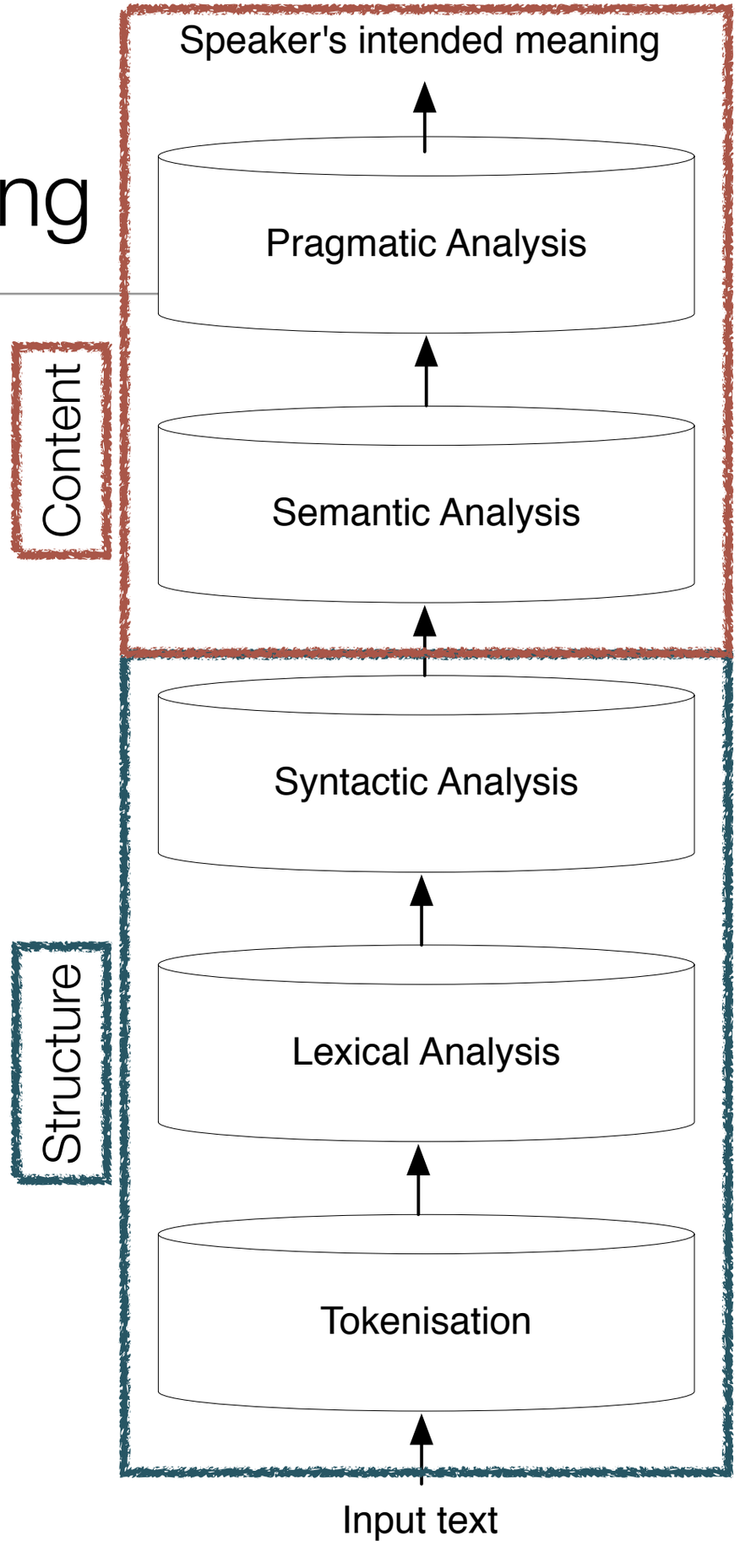
Source, path, goal,

Pragmatics

- In real live language is stretched to serve a purpose in a context (*people always try to make sense*)
    - The three pizzas still need to pay. (**Metonymy**)
    - The salty peanut fell in love with the cashew nut. (**N400** effect in the brain)
    - Can you close the window, please? (Form: question, Meaning: Request)
    - It is a bit cold here, isn’t it? (Form: a declarative sentence, Meaning: request)
- Further readings
    - Nieuwland, Mante S., and Jos JA Van Berkum. "When peanuts fall in love: N400 evidence for the power of discourse." *Journal of cognitive neuroscience* 18, no. 7 (2006): 1098-1111.
    - Grice, H. Paul, Peter Cole, and Jerry L. Morgan. "Logic and conversation." *1975* (1975): 41-58. => ***maxims of conversation***
    - Searle, John R.. ***Speech acts***: An essay in the philosophy of language. Vol. 626. Cambridge university press, 1969.

Part II: Natural Language Processing

- Complex problem is broken down into a number of smaller problems
- Simple, structural problems solved first and higher-level semantic tasks are solved later, using the output of earlier modules as input:
  - pipeline architecture with dependencies across modules
- For each problem different techniques:
  - knowledge-base & rules (linguistic knowledge)
  - machine learning (supervised and unsupervised), data driven



We always need to do preprocessing

- First problem, what is a word, what is a sentence?

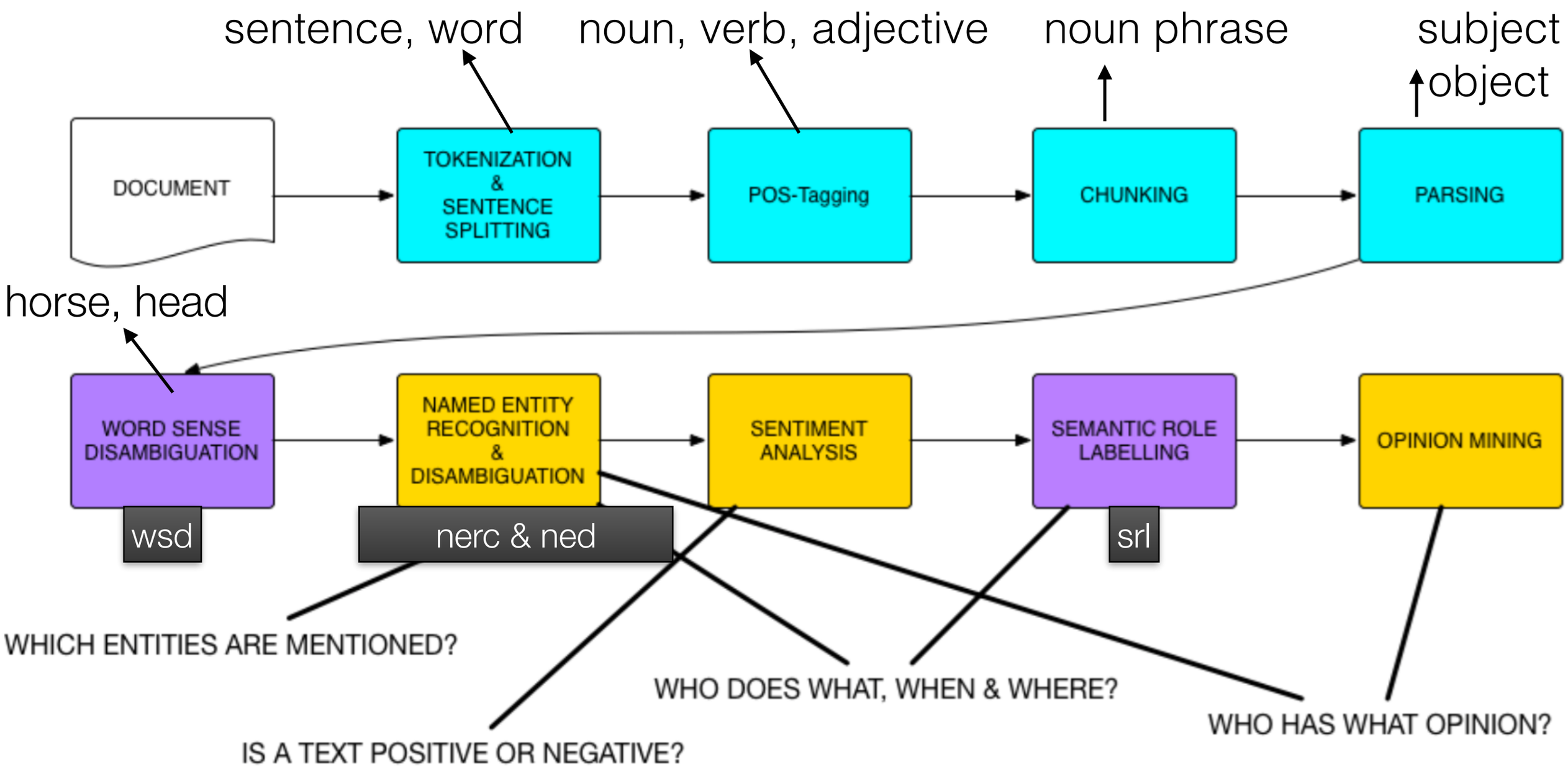
- **Tokenization**

    - nitty-gritty, data-base, (semi-)irony, $523,45, 21st century, 9-11, Encoding issues (Latin, UTF-8/16, diacritiçćčs, “‘’`` qoutes…), don’t, men’s, end-of-sentence hyphens

- **Sentence splitting**

    - Dr., Mrs., Bol.com, 7.5, etc. etc, white spaces, TABs, new lines, HTML markup <body><h1></h1></body>

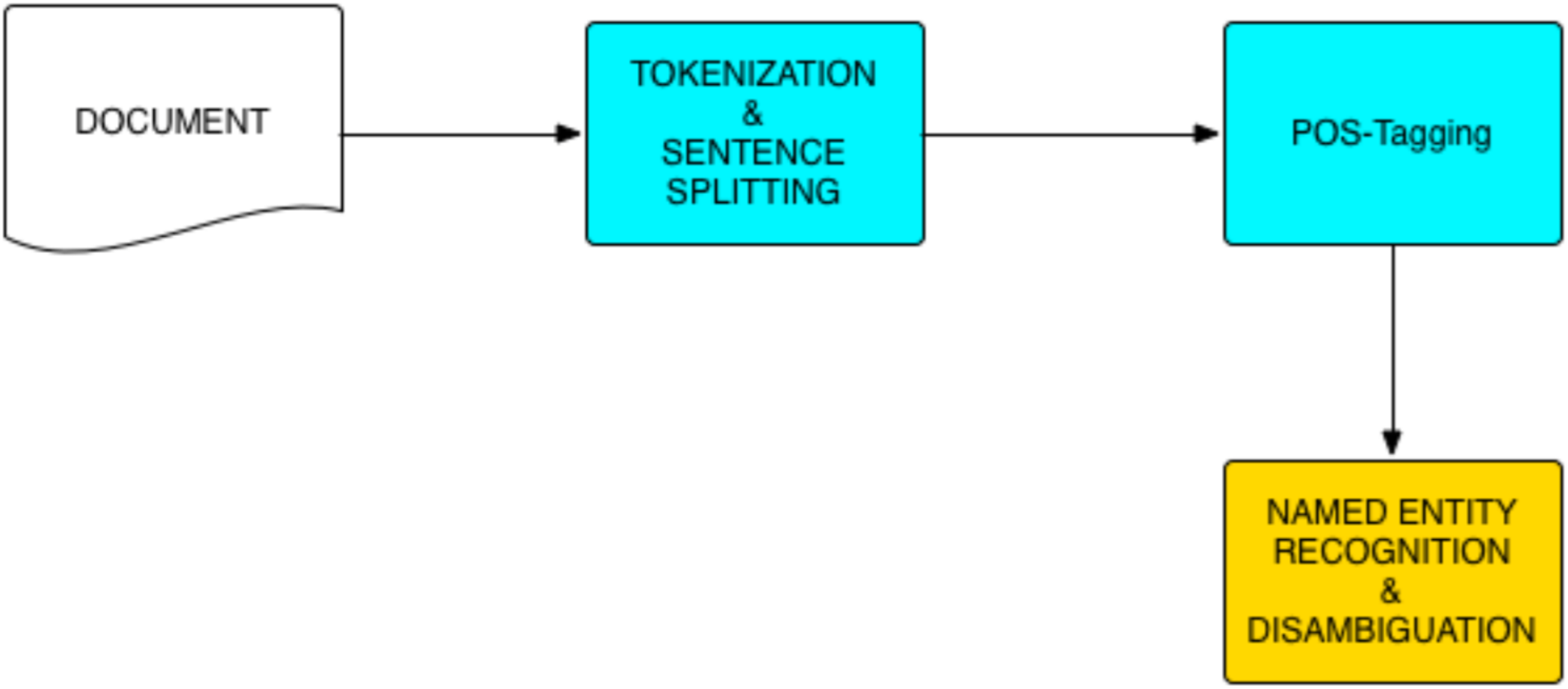
Example of an NLP pipeline



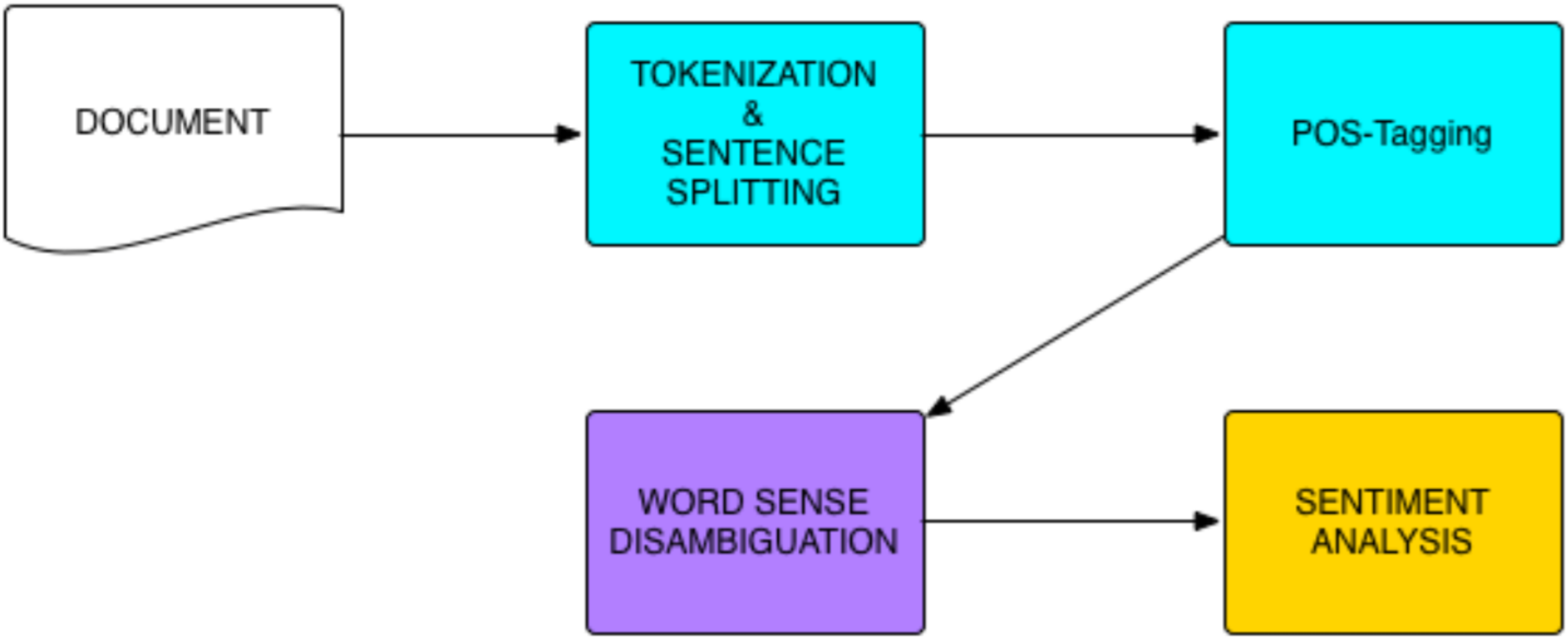
Communication between modules

- Modules in a pipeline often reuse information from a previous step
- Modules must therefore be able to communicate with each other
- Frameworks that support integration of several steps
- Representation formats that are compatible or easily adaptable

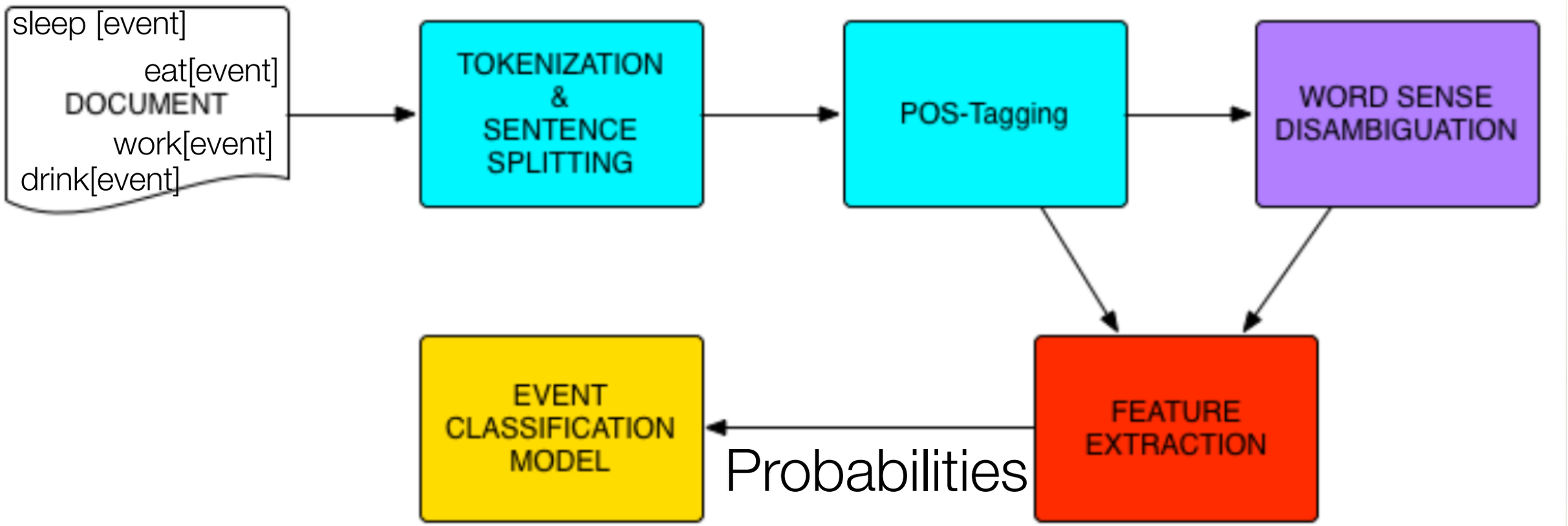
Named Entity Recognition Pipeline Example



Sentiment Analysis Pipeline Example



Example of Supervised Machine Learning

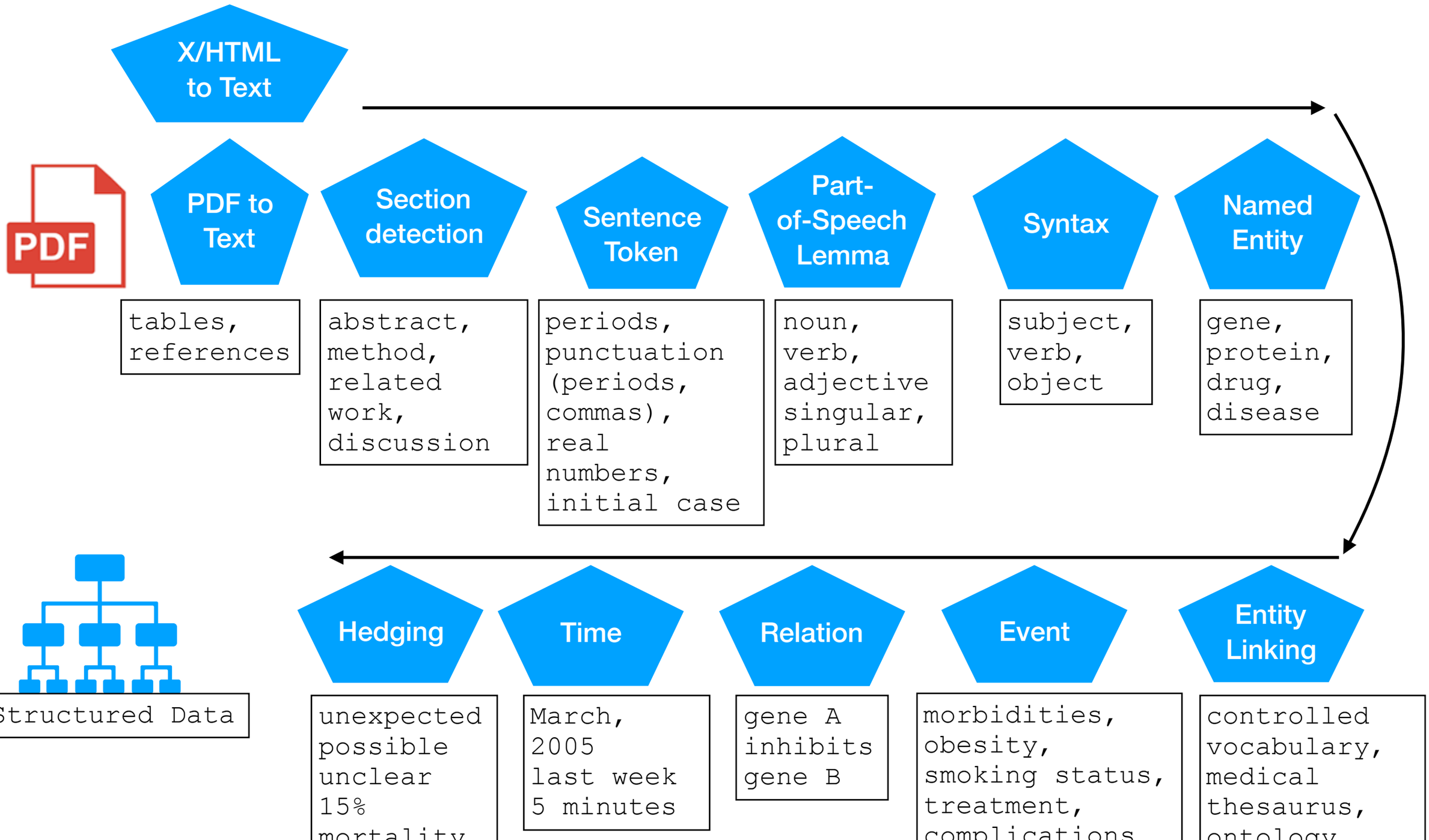


the/a, D + W event, V: 1%
the/a, D + W event, N: 30%
*, N + W event, V: 80%

surrounding words
lemmas, meanings
part-of-speech

Medical NLP Pipeline

scientific articles & clinical documents: admission notes, discharge summaries, radiology reports, pathology reports, etc.



NLP pipeline plumbing

- Bottom-up processing from low-level tokens to high-level interpretation.

- Single isolated tasks: tokenization, pos-tagging, parsing, named-entity-detection (NERC, NED), word-sense-disambiguation (WSD), semantic-role-labeling (SRL), time expressions.

- Pass output from one module to the next, no way back.

- Use simple left/right contexts of target tokens.

- Consider each token in isolation, one solution does not know about the solution to other tokens of the same type.

Errors in Word-Sense-Disambiguation for monosemous words due to part-of-speech errors

Competition	Monosemous	Wrong	Examples
Senseval2	499 (20.9%)	37.5%	gene.n (*suppressor_gene.n*), chance.a (*chance.n*) next.r (*next.a*)
Senseval3	334 (16.6%)	44.1%	Datum.n (*data.n*) making.n (*make.v*) out_of_sight (*sight*)
Semeval2007	25 (5.5%)	11.1%	get_stuck.v, lack.v, write_about.v
Semeval2010	31 (2.2%)	97.9%	Tidal_zone.n pine_marten.n roe_deer.n cordgrass.n
Semeval2013 (lemmas)	348 (21.1%)	1.9%	Private_enterprise, developing_country, narrow_margin

Performance
in relation to domain specificity and complexity

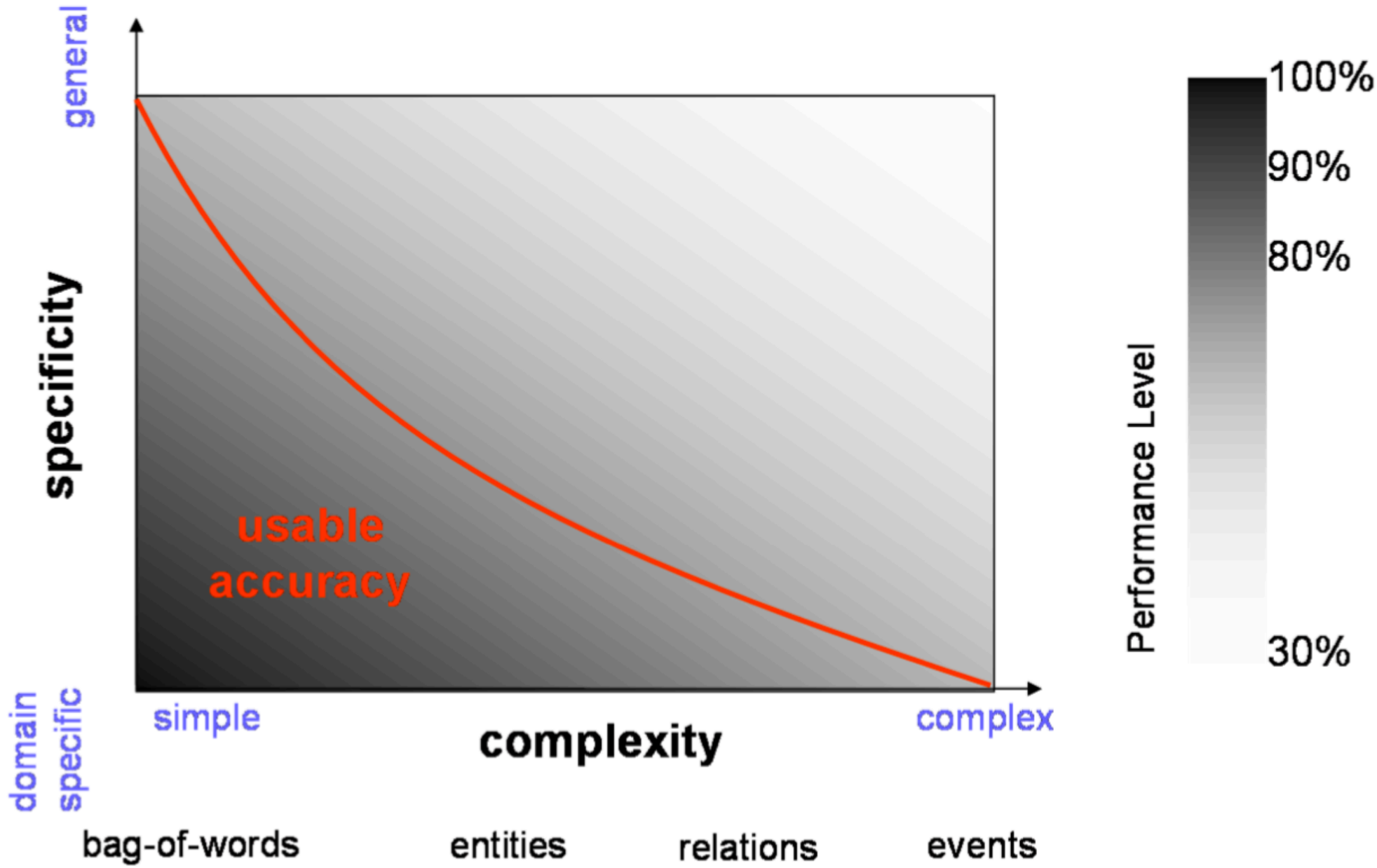


Figure 1.3: Performance Tradeoffs for NLP tasks

Recap Part II: NLP Pipelines

- NLP tasks often involve more than one analysis
- NLP pipelines are sequences of NLP modules carrying out linguistic analyses
- Some NLP modules need linguistic analyses of other modules as input: i.e. they depend on previous analyses carried out by other modules
- NLP modules need to be interoperable
- Performance of individual modules cannot simply be averaged over the pipeline due to error propagations

NLP Documentation

Computational linguistics

- Juravsky and Martin, Speech and Language Processing, 3rd edition:
  - 2017 edition: JuravskyMartin_ed3book-2017.pdf
  - Current edition https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit: www.nltk.org/book/

Programming

- Anaconda Python distribution: https://www.anaconda.com/distribution/
- Python documentation: https://docs.python.org/3/.
- Code editors :Atom (https://atom.io/) and PyCharm (https://www.jetbrains.com/pycharm/) .
- References:
  - Python cookbook: chimera.labs.oreilly.com/books/1230000000393/index.html
  - Think Python: greenteapress.com/wp/think-python-2e/

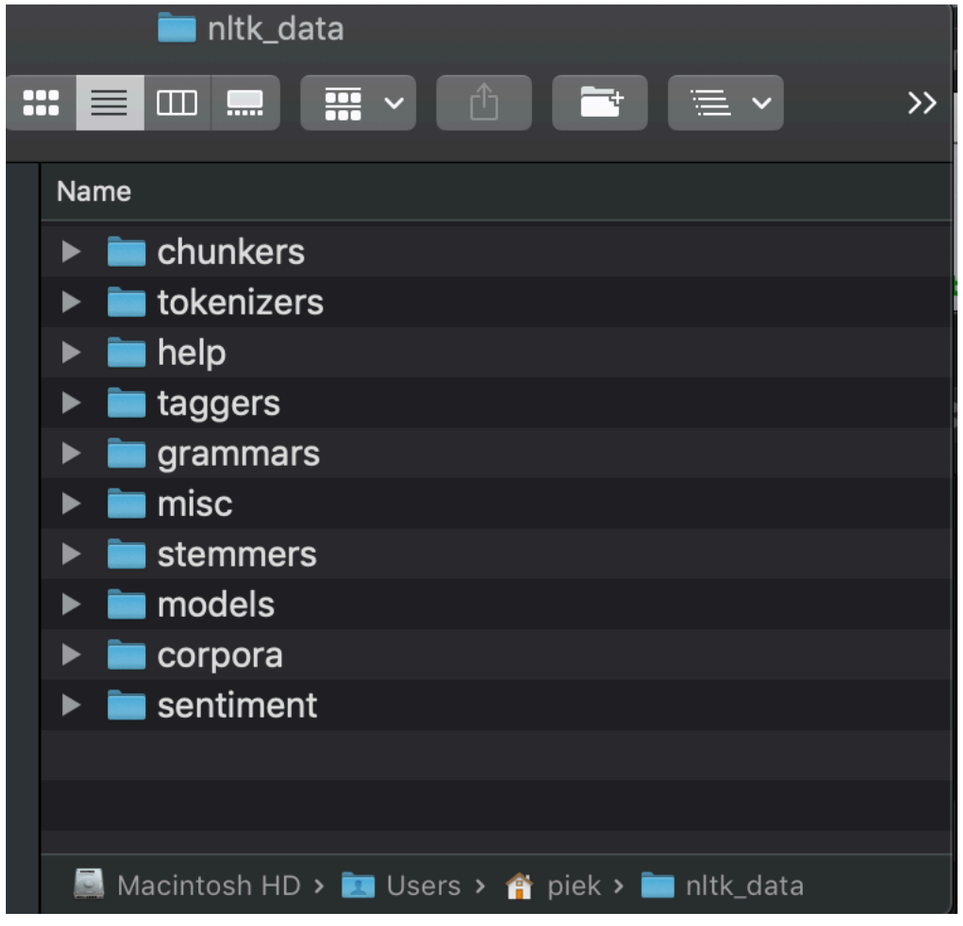
Mailing lists

Subscribe to get information about conferences, jobs, etc.

- linguistlist.org/
- https://mailman.uib.no/listinfo/corpora
- https://lists.uni-duesseldorf.de/mailman/listinfo/semantik
- www.siggen.org/mailing.html

Linguistic processors

- NLTK, SpaCy package many linguistic processors for many languages
    - Sentence segmentation
    - Tokenisation
    - Lemmatisation
    - Part-of-Speech tagging
    - Chunkers, constituency parsers, dependency parsers
- For many tasks these processors are called before performing text mining
- Specific processing is needed for micro-blogs:
    - http://www.cs.cmu.edu/~ark/TweetNLP/



Some NLP for Dutch

- STEVIN: https://ivdnt.org/taalmaterialen and CLARIAH: https://www.clariah.nl/en/
- Morpho-syntactic processing:
    - Frog: https://languagemachines.github.io/frog/
    - Alpino: https://www.let.rug.nl/vannoord/alp/Alpino/
- Semantic processing of text:
    - VU-reading-machine
        - Docker image https://cloud.docker.com/u/vucltl/repository/docker/vucltl/vu-rm-pip3
        - Source code: https://github.com/cltl/vu-rm-pip3
- Resources:
    - E-Lex: https://ivdnt.org/downloads/taalmaterialen/tstc-e-lex
    - Dutch WordNet (http://wordpress.let.vupr.nl/odwn/) and FrameNet (http://dutchframenet.nl)
    - ANS, Algemene Nederlandse Spraakkunst: http://www.let.ru.nl/ans/

More details

Part III: Language as data

- How many words exist in a language and how many do you use?
  - standard dictionary: 100K+ words
  - adults: 20K (active) - 40K (passive)
  - 1 year news: millions of unique words, every day 3 new words
  - ***BUT!*** 3000 words (5% of the vocabulary) covers 95% of the news!!!
- How much does a person read?
  - 2,000 per day (4 pages), 1 million per year…

More details

What is corpus linguistics

- Corpus linguistics is the discipline that defines methods to analyse collections of text to test hypotheses on language
- Goal:
    - extend, test and redefine linguistic theories
    - derive new linguistic theories
- What is a theory?
    - Set of rules that make predictions: e.g. what is a grammatical sentence, how does language change?
- Lexicon: which words are stored with which properties in our brain
- Train computer models that ‘fit the data’: speech recognition, dialogue systems, machine translation

More details

Types of analyses

- Qualitative analysis (***deep bit few data***):
  - dingen die kunnen worden geobserveerd maar niet of moeilijk gemeten: *de waardering van rechtsinstituties in de Nederlandse samenleving in de loop der eeuwen*
  - inventariseert karakteristieken van het geanalyseerde: wat is een *rechtsinstituut*, waaruit blijkt *waardering*, *wie* drukt welke waardering uit, wat is de *autoriteit* van die personen?
- Quantitative analysis (***superficial statistics but a lot of data***):
  - analyse van dingen die gemeten/geteld kunnen worden, bv. hoe vaak (***eenduidige)*** karakteristieken voorkomen. Bv een analyse van het sentiment in alle teksten waarin het woord “politie” (“agent”) genoemd wordt.
  - hoeveel moet je meten/tellen om tot significante conclusies te komen?

More details

Aspecten opzet corpuslinguïstische analyse

- Communicatiemodus: gesproken, geschreven, video
- Dataverzameling
- Wel of niet geannoteerd?
- Dataselectie
- Meertalig of monolinguaal?

Dataverzameling

Token = woordvoorkomen, Type = woordvorm
1 T= 1 token, 1T =1 miljoen tokens

- Hoe zorg **je** ervoor dat **je** data sample representatief is voor de taal/het linguïstische fenomeen dat **je** wilt onderzoeken? *[20 tokens, 17 types]*
    - Monitor corpus: groeit over tijd en is vooral representatief door zijn ***grootte*** bv Bank of English (BoE, 650 MT)), Corpus of Contemporary American English (COCA 500MT) of het WWW (corpus in het wild)
    - Sample/Gebalanceerd corpus: ***gebalanceerd*** en representatief voor een steekproef uit een bepaalde ***populatie*** bv Lancaster/Oslo/Bergen (LOB, 1MT), Brown corpus (1MT), British National Corpus (BNC, 100MT)
        - Speciale corpora: Child Language Data Exchange System (CHILDES, gesproken, diverse talen)
- Corpora list: http://clu.uni.no/icame/corpora/sites.html
- Nederlands: http://taalunieversum.org/inhoud/corpora
    - Corpus Geschreven Nederlands (SoNaR (500MT)), Corpus Gesproken Nederlands (CGN, 1MT)),

More details

Het Internet als een corpus

- Google
- http://www.webcorp.org.uk/live/index.jsp
- http://wse1.webcorp.org.uk

More details

Balans, Representativiteit en Vergelijkbaarheid

- Hoe weet je of je echt alle dimensies van de taalvariatie hebt afgedekt?
- In sommige gevallen is er niet veel data (e.g. ‘dialecten’, ’bedreigde talen’)
- DutchSemCor project: http://wordpress.let.vupr.nl/
- Zoek 25 voorbeelden van alle betekenissen van woorden zoals “paard”, “band”, “spelen”, “lopen”
- 23% van de betekenissen geen voorbeeld in SoNaR corpus van 500 miljoen woorden
- Wat zijn de betekenissen van “band”?
  - “Ik was op vakantie in Band.”

Band

Band kan verwijzen naar:

Materiaal [bewerken]

Strook [bewerken]

- Ligament (anatomie), een band van bindweefsel om een gewricht
- Band (bouwkundig), een horizontale versiering in een gevel
- Omslag om een boek
  - Een boek met de omslag, zie Volume (fysieke informatiedrager)
- Chromosoomband, een gebied op een chromosoom
- Lopende band, een continu voortbewegende strook materiaal om goederen mee te trans
- Magneetband, een strook folie bedekt met een magnetische laag, gebruikt om geluid, vi
- Obi (zelfverdediging), een kledingband gedragen bij Japanse zelfverdedigingskunsten e

Om een wiel [bewerken]

- Luchtband, een opblaasbare wielband, het meest gebruikt bij wegvoertuigen
- Massieve band, een metalen wielband om een houten wiel of treinwiel of een rubberban

Overdrachtelijk [bewerken]

- Band (radio), een deel van het frequentiespectrum bestemd voor radioverkeer
- Band (samenleving), de eenvoudigste vorm van menselijke sociale organisatie
- Frequentieband, aaneengesloten bereik van frequenties van licht, radiogolven, straling o

Muziekgroep [bewerken]

- Muziekgroep of popgroep, de algemene benaming voor een groep muzikanten
- Jazzband, een muziekgezelschap dat jazz speelt
- The Band, een Canadees-Amerikaanse rockband uit de jaren zestig en zeventig

Film en televisie [bewerken]

- Band (documentaire), een documentaire uit 1998 van Duane Condor
- De Band, een Nederlandse komedieserie

Plaats [bewerken]

- Bánd, een plaats in Hongarije
- Band (Mureș), een plaats in Roemenië

Hoofdpagina
Vind een artikel
Vandaag
Etalage
Categorieën
Recente wijzigingen
Nieuwe artikelen
Willekeurige pagina

Informatie
Gebruikersportaal
Snelcursus
Hulp en contact
Donaties

Hulpmiddelen
Links naar deze pagina
Verwante wijzigingen
Bestand uploaden
Speciale pagina's
Permanente koppeling
Paginagegevens
Wikidata-item
Deze pagina citeren

Afdrukken/exporte
Boek maken
Downloaden als PDF
Printvriendelijke versie

In andere talen

Betekenis ' band '

Je hebt gezocht op het woord: band.

[1]band *(de; m;* meervoud: *banden)*

1 reep van stof die dient om te binden

2 iets dat bindt: *iem., iets aan banden leggen* beteugelen; *uit de band springen* zich laten gaan

3 met lucht gevulde ring van rubber om een wiel

4 omslag met een sterke rug waarin een boek wordt gebonden

5 binnenrand van een biljart

6 transportband: *de lopende band* systeem waarbij op een transportband een voorwerp verschillende arbeiders passeert, die elk een bep. handeling verrichten; *aan de lopende band* telkens, heel vaak

7 strook magnetisch materiaal om beeld en geluid vast te leggen

¶ *(bij vechtsporten) de groene, zwarte enz. band* bij wijze van sterkteaanduiding gedragen

[2]band *(het; o)*

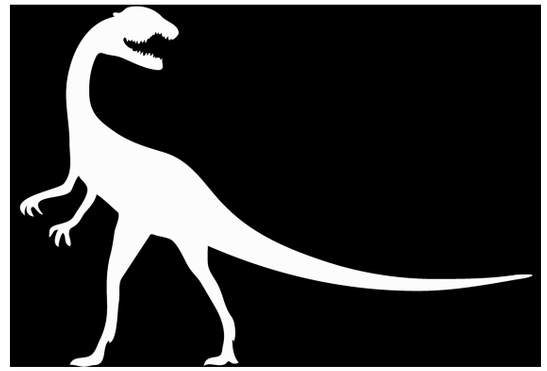
1 lintvormig weefsel

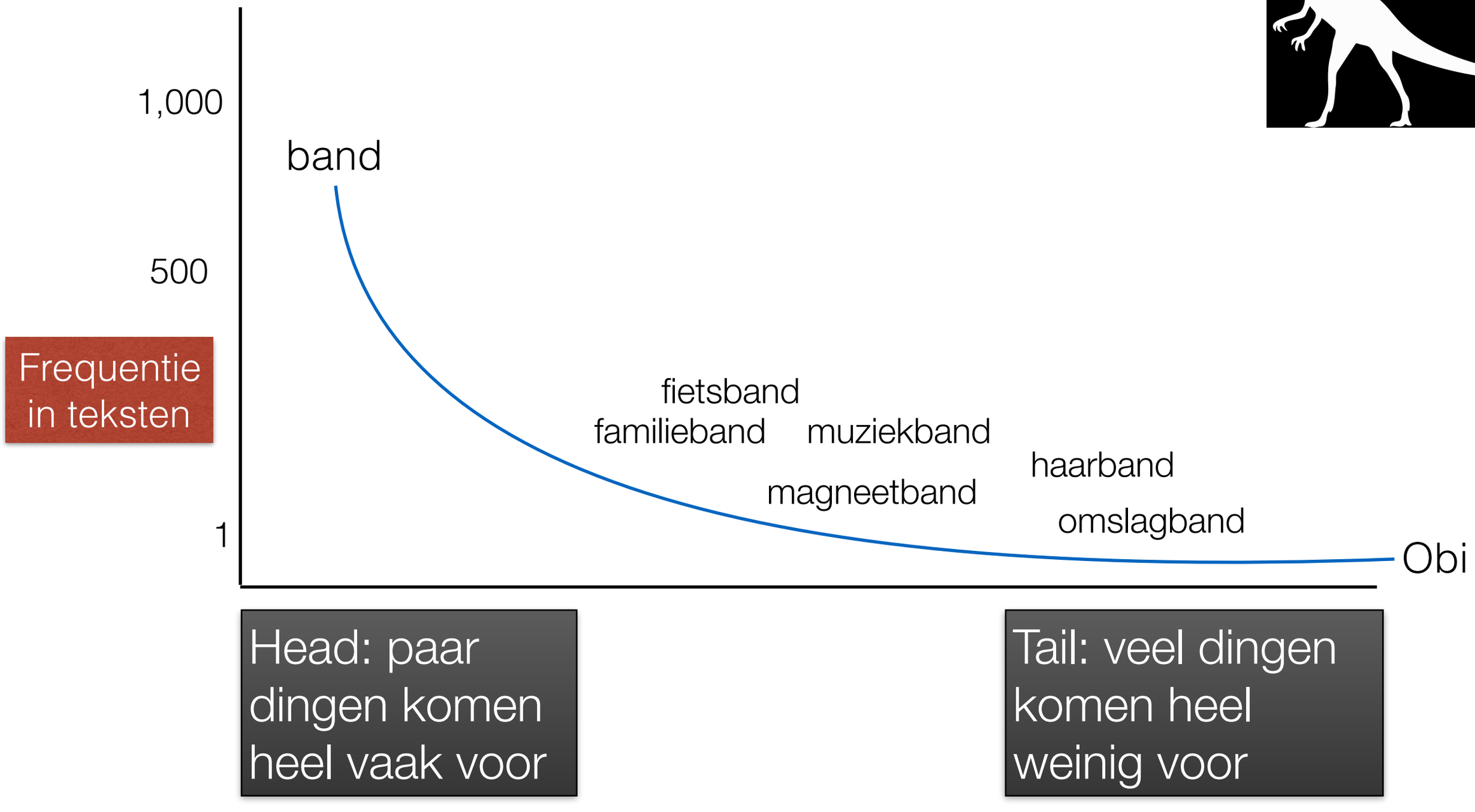
[3]band *(de; m;* meervoud: *bands)*

1 (muziek)korps; militaire kapel; = popgroep

More details

LONGTAIL DETAILS





5% woorden, 2% van de betekenissen, 95% van het nieuws

More details

Andere vragen?

- Woorden en betekenissen op sociale media?
- Verschillende groepen mensen? Leeftijd, geslacht, opleiding, regio, 1st/2de taal, etc.
- Veranderingen over de tijd
- Registers en context: thuis, school ,werk, kroeg, sportveld
- meer dan woorden: uitdrukkingen, zinnen, paragraaf …..
- minder dan woorden: syllabes, vervoeging, afleiding, samenstelling, uitspraak, …
- meer dan tekst: video, multimedia web pagina’s, plaatjes en taal, geluiden en geuren en taal….

More details

Wel of niet geannoteerd?

- Voor sommige analyses is de tekst op zichzelf niet genoeg.
    - Annotaties zoals grammaticale categorie, syntactische structuur, uitspraak, betekenis etc kunnen zinvol zijn
        - Bij grote corpora moeten ze (semi-)automatisch toegevoegd worden
- Annotaties kunnen ‘inline’ of ‘stand-off’ zijn (en daarbinnen zijn weer veel variaties mogelijk). Dit komt interoperabiliteit niet ten goede.

More details

Total accountability en dataselectie

- Computer analyse zou je in staat moeten stellen om niet een specifiek “sample” te onderzoeken maar het gehele corpus -> kwantitatief onderzoek
    - Maar het corpus zelf is natuurlijk ook een sample
    - Voor analyse van sommige corpora is veel rekenkracht nodig
- Dataselectie: soms kan een tegenvoorbeeld voldoende zijn om een theorie te verwerpen (falsificeerbaarheid)

Meertalig of monolinguaal?
Een genre of alle genres? Nu, vroeger, diachroon?

More details

- Deze keuzes hangen af van je onderzoeksvraag
- Verschillende types meertalige corpora:
    - Type A (translation corpus): bronteksten + vertalingen in een of meerdere andere talen
    - Type B (parellel corpus): Paren of groepen monolinguale corpora die volgens hetzelfde corpusdesign zijn samengesteld
    - Type C: Een combinatie van A & B.