# DD2418 Language Engineering
# Lecture 1a: Intro

Johan Boye, KTH

March 16, 2020

# What this course is about

This course is about processing human languages.

- English, Swedish, Chinese, Russian, Arabic, ...

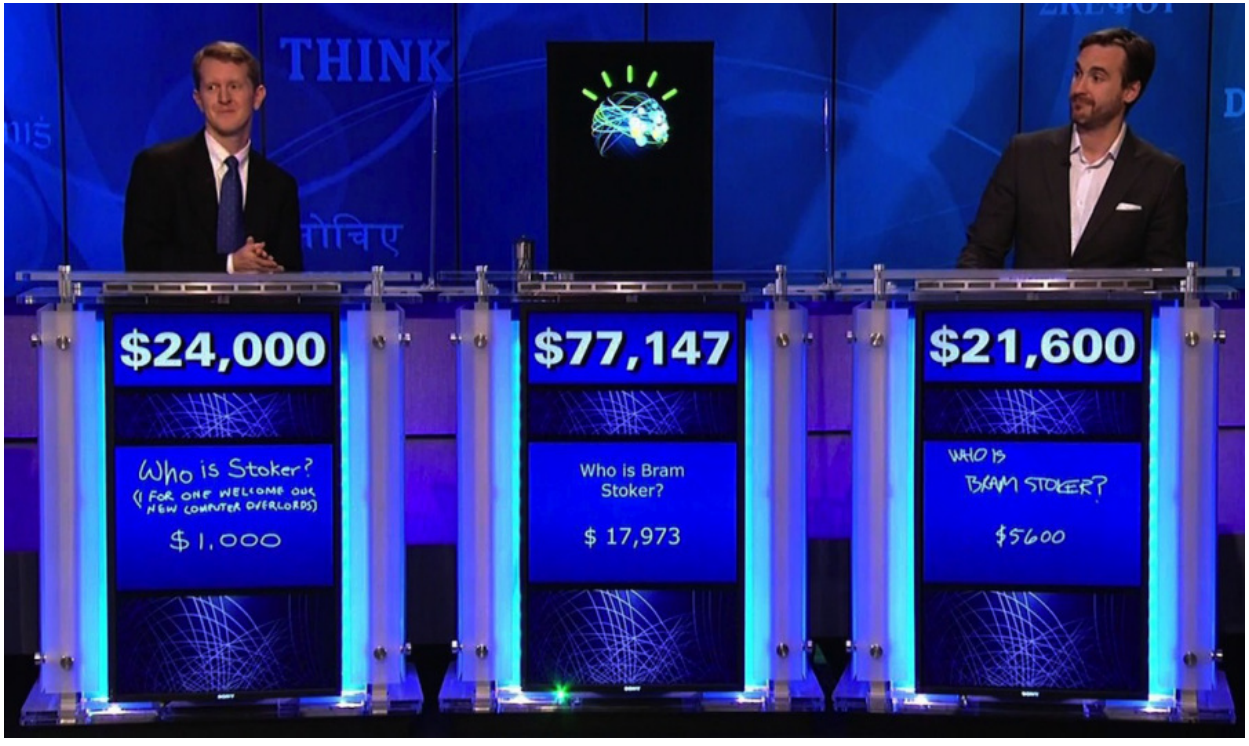
What is “processing”?

- One-liner: “Making computers understand language”
- More precisely: Extracting useful information from natural language, and generating (correct, useful) language.

# What's the point?

- A lot of human knowledge is represented as text.
- Lots of useful applications.
- Understanding language is a key aspect in the AI programme.
- We can gain insights about human language from the study of computational models.
- Language processing is one of the main drivers of machine learning (together with image recognition).
- Getting an interesting and well-paying job!

# Question answering: Watson




On February 16, 2011, IBM's computer system Watson defeated the world's best human Jeopardy champions

https://vimeo.com/222234104

# Language engineering applications

- Question-answering
- Speech recognition
- Machine translation
- Search engines
- Natural-language interfaces
- Text summarization
- Text classification
- Word prediction
- Spell checking
- Grammar checking
- ...

# Some things we don't know about language

- How old is it? (dunno, but perhaps around 100,000 years).
- How did it arise?
- Why do children learn it so effortlessly?
- How does the brain generate and analyse it?

# Some things we do know about language

- It's uniquely human.
- There are 5,000-7,000 known languages.
- The relation between words and meaning is arbitrary.
- Language is rule-bound.
- Language is productive.

# ’Arbitrariness of the sign’

- The relation between words and meaning is arbitrary.
  - The word *dog* doesn’t look like a dog or bark like a dog...
  - ... but means ’dog’ all the same.
  - Seems obvious but is an important prerequisite for the effectiveness of language.
- This observation is usually attributed to *Ferdinand de Saussure* (1857-1913).

# Language is rule-bound and productive

Language can be broken down to smallest units (words), which are combined using the rules of the language.

These rules are a *naturally occurring phenomenon*, not something we learn in school.

Using these sub-conscious rules, we can produce and understand an infinite number of sentences, e.g.:

- He went skiing with a kangaroo and a watermelon in his left shoe.

Even if words are arbitrary, language structure definitely is not!

# Processing language by computers (1)

Rule-based approaches

- The earliest work strived to uncover the subconscious rules that guide our language understanding.
- Grammars, automata, logic formulas, etc.
- Great when you don't have much data
- Predominant approach beginnings-1990's
- Rationalism

# Processing language by computers (2)

Statistical / “traditional” machine-learning approaches

- Learns patterns directly from text data (from *corpora*)
- Great if you have lots of data
- Predominant approach late 1990’s-2014 (about)
- Empiricism

Neural approaches

- Great if you have huge amounts of data
- End-to-end systems

# Why is language understanding hard?

Language is *underspecified*

Language interpretation *requires knowledge about the world*.

Language is *ambiguous*

- “John made the pie in the fridge”

+ many more (smaller) challenges...

# Lectures

Provisional lecture schedule:

1. Intro. Words.
2. Syntax and parsing
3. Statistical language models
4. Spell checking and grammar checking (Viggo Kann)
5. Text classification
6. Logistic regression
7. Word embeddings
8. Neural networks
9. Language processing with recurrent neural networks (RNNs)

# DD2418 Language Engineering
# Lecture 1b: Words

Johan Boye, KTH

# Levels of linguistic analysis

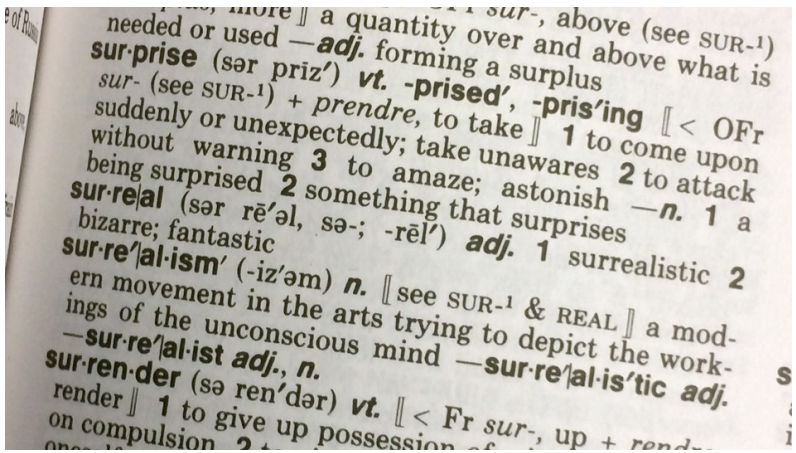
| | |
|---|---|
| Words | Morphology, Phonology |
| Sentences | Syntax |
| Meaning | Semantics |
| Language use | Pragmatics |

# Some terminology

A *corpus* (pl. *corpora*) is a collection of texts, possibly annotated.

A *lexicon* or *vocabulary* is a list of all the unique words in a corpus.

The *lemma* (pl. *lemmata*) of a word is what you look up in a dictionary.



# Words and tokens

*A hat and a hat make two hats*

contain 8 *tokens*, but 5 words (not counting morphological variants), or 6 words (counting morphological variants).

In language engineering, we see two variants:

full-form lexicon Save all forms of every word in the lexicon (*go, goes, gone, went, going,* etc.)

morphological analysis Store only lemmas + rules to generate all the other forms.

# Word classes

Words can be divided into classes depending on their use in the language.

- Noun, verb, adjective, adverb, preposition, pronoun, conjunction, interjection, determiner, etc.
- These classes are often called *parts-of-speech* (or *POS tags*).

Lots of debate in linguistics about their nature and generality.

The idea of word classes can be traced back to Aristotle (384-322 BC) and Dionysius Thrax (about 170-90 BC).

# Quiz

Determine the word classes in the following sentence:

I should go home today

# Quiz

Determine the word classes in the following sentence:

| I | should | go | home | today |
|---|---|---|---|---|
| Pronoun | Auxiliary verb | Verb | Adverb | Adverb |

# Open and closed classes

Closed classes, e.g. :

- determiners (a, an, the, some, . . .)
- pronouns (she, her, I, you, me, . . .)
- prepositions (on, to, under, from, . . .)

Open classes

- nouns, verbs, adjectives, adverbs

# SUC (Stockholm-Umea Corpus) tag set

| Code | Swedish category | Example | English translation |
|---|---|---|---|
| | | | |
| AB | Adverb | *inte* | Adverb |
| DT | Determinerare | *denna* | Determiner |
| HA | Frågande/relativt adverb | *när* | Interrogative/Relative Adverb |
| HD | Frågande/relativ determinerare | *vilken* | Interrogative/Relative Determiner |
| HP | Frågande/relativt pronomen | *som* | Interrogative/Relative Pronoun |
| HS | Frågande/relativt possessivt pronomen | *vars* | Interrogative/Relative Possessive |
| IE | Infinitivmärke | *att* | Infinitive Marker |
| IN | Interjektion | *ja* | Interjection |
| JJ | Adjektiv | *glad* | Adjective |
| KN | Konjunktion | *och* | Conjunction |
| NN | Substantiv | *pudding* | Noun |
| PC | Particip | *utsänd* | Participle |
| PL | Partikel | *ut* | Particle |
| PM | Egennamn | *Mats* | Proper Noun |
| PN | Pronomen | *hon* | Pronoun |
| PP | Preposition | *av* | Preposition |
| PS | Possessivt pronomen | *hennes* | Possessive |
| RG | Grundtal | *tre* | Cardinal number |
| RO | Ordningstal | *tredje* | Ordinal number |
| SN | Subjunktion | *att* | Subjunction |
| UO | Utländskt ord | *the* | Foreign Word |
| VB | Verb | *kasta* | Verb |

# Penn Treebank tag set

| Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|
| CC | coordin. conjunction | *and, but, or* | SYM | symbol | +,%, & |
| CD | cardinal number | *one, two* | TO | "to" | *to* |
| DT | determiner | *a, the* | UH | interjection | *ah, oops* |
| EX | existential 'there' | *there* | VB | verb base form | *eat* |
| FW | foreign word | *mea culpa* | VBD | verb past tense | *ate* |
| IN | preposition/sub-conj | *of, in, by* | VBG | verb gerund | *eating* |
| JJ | adjective | *yellow* | VBN | verb past participle | *eaten* |
| JJR | adj., comparative | *bigger* | VBP | verb non-3sg pres | *eat* |
| JJS | adj., superlative | *wildest* | VBZ | verb 3sg pres | *eats* |
| LS | list item marker | *1, 2, One* | WDT | wh-determiner | *which, that* |
| MD | modal | *can, should* | WP | wh-pronoun | *what, who* |
| NN | noun, sing. or mass | *llama* | WP$ | possessive wh- | *whose* |
| NNS | noun, plural | *llamas* | WRB | wh-adverb | *how, where* |
| NNP | proper noun, sing. | *IBM* | $ | dollar sign | *$* |
| NNPS | proper noun, plural | *Carolinas* | # | pound sign | *#* |
| PDT | predeterminer | *all, both* | " | left quote | ' or " |
| POS | possessive ending | *'s* | " | right quote | ' or " |
| PRP | personal pronoun | *I, you, he* | ( | left parenthesis | [, (, {, < |
| PRP$ | possessive pronoun | *your, one's* | ) | right parenthesis | ], ), }, > |
| RB | adverb | *quickly, never* | , | comma | , |
| RBR | adverb, comparative | *faster* | . | sentence-final punc | . ! ? |
| RBS | adverb, superlative | *fastest* | : | mid-sentence punc | : ; ... – - |
| RP | particle | *up, off* | | | |

# Universal Dependencies tag set

Universal Dependencies is an initiative to create a multilingual tagset and a set of multilingual analysis tools.

| Open class words | Closed class words | Other |
|---|---|---|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

# Part-of-Speech tagging

The *part-of-speech tagging* problem is to assign a POS tag to each word in a text.

Why is this useful?

- Speech synthesis (**in**sult (noun) vs in**sult** (verb))
- Syntactic analysis ("*Time flies like an arrow*")
- Finding content words in text

  ADJ NOUN "*linear function*"
  NOUN NOUN "*regression coefficient*"
  ADJ ADJ NOUN "*Gaussian random variable*"
  NOUN PREP NOUN "*degrees of freedom*"
- Back-off model for machine learning (when we have sparse data).

# Ambiguities

Some words can belong to more than one class, e.g. *like*:

| | |
|---|---|
| VERB | "*I like her.*" |
| NOUN | "*He got a like on Facebook.*" |
| ADJ | "*The portrait is very like.*" |
| ADV | "*This is, like, crazy!*" |
| ADP | "*It looks like an accident*" |
| SCONJ | "*He acted like he was all alone*" |

# Ambiguities

Swedish also has this kind of words, e.g. *så*:

ADV “*Så gick det till.*”

PRON “*På så sätt!*”

SCONJ “*Han åt så han blev mätt*.”

INTJ “*Så, det var intressant att höra!*

VERB “*Man måste så innan man kan skörda.*”

NOUN “*Grisarna drack ur en så.*”

# New words

Even with a large dictionary, you will always encounter new words.

- Names
- Neologisms like *metrosexual*, to *google*, etc.
- New concepts (*bogvisir*)

# Part-of-speech tagging

Retrieve all possible tags from a dictionary, then decide which ones are the most likely, e.g. :

| *I* | *like* | *plays* | *about* | *bolliwogs* |
|---|---|---|---|---|
| PRON | VERB/NOUN ADJ/ADV /ADP/CONJ | VERBB/NOUN | ADV/ADP/ADJ | ? |

| PRON | VB | NOUN | PREP | NOUN |
|---|---|---|---|---|

# Difficulty of POS tagging

POS tagging is considered to be an “easy” problem.

State-of-the-art POS taggers have about 97% accuracy, which is as good as it gets (what human experts can agree on, the *gold standard*.

However, a straightforward method (a *baseline*) already gives about 90% accuracy.

- Tag every word with its most frequent tag.
- Tag unknown words as nouns.

The high accuracy is partly due to:

- Many words are unambigous (including *a*, *an*, *the*).
- Punctuation (“.”, “!”, “?”) is unambigous.

# Difficulty of POS tagging

Some POS tagging decisions are difficult for computers:

ADP “*He turned off the road*”

PART “*He turned off the radio*”

# DD2418 Language Engineering
# Lecture 1c: The structure of words

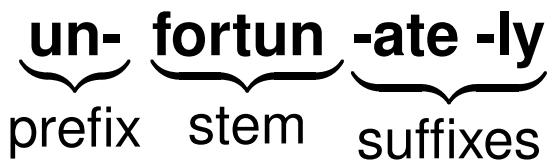
Johan Boye, KTH

# The structure of words

*Morphology* is the study of how *words* are built from smaller units called *morphemes* (smallest *meaningful* unit)

Two kinds of morphemes:

| | |
|---|---|
| stem | the core unit |
| affixes | small units signalling various grammatical functions |




Note that the stem doesn't have to be a word! Stem $\neq$ lemma!

# Affixes

Affixes come in four varieties:

| | |
|---|---|
| prefix | *un*-familiar |
| suffix | quick-*ly* |
| infix | (*sv.*) korru-*m*-pera |
| circumfix | (*ge.*) *ge*-sag-*t* |

# Word formation

Words can be formed from other words by:

inflection (*sv.* böjning) *cat - cats*

derivation (*sv.* avledning) *friend - friendly - friendliness*

compounding (*sv.* sammansättning) *smartphone, anti-missile*

# Inflections: Verbs

A typical English verb has 4 or 5 forms.

- *ask - asks - asked - asking*
- *go - goes - gone - went - going*

Swedish: about 10 forms

- *äta - äter - åt - ätit - ätande - ät - ätas - äts - åts - ätits - äten*

French: >40 forms

Classic Greek: 350 forms*

Turkish: 3 million forms*

*S. Pinker (1997) *The language instinct*, Penguin.

# Inflections: Verbs

## *Parler*

**The verb *parler* "to speak", in French orthography and IPA transcription**

| | Indicative | | | | Subjunctive | | Conditional | Imperative |
|---|---|---|---|---|---|---|---|---|
| | **Present** | **Simple past** | **Imperfect** | **Simple future** | **Present** | **Imperfect** | **Present** | **Present** |
| **je** | parl-e /paʁl/ | parl-ai /paʁle/ | parl-ais /paʁlɛ/ | parl-erai /paʁləʁe/ | parl-e /paʁl/ | parl-asse /paʁlas/ | parl-erais /paʁləʁɛ/ | |
| **tu** | parl-es /paʁl/ | parl-as /paʁla/ | parl-ais /paʁlɛ/ | parl-eras /paʁləʁa/ | parl-es /paʁl/ | parl-asses /paʁlas/ | parl-erais /paʁləʁɛ/ | parl-e /paʁl/ |
| **il** | parl-e /paʁl/ | parl-a /paʁla/ | parl-ait /paʁlɛ/ | parl-era /paʁləʁa/ | parl-e /paʁl/ | parl-ât /paʁlɑ/ | parl-erait /paʁləʁɛ/ | |
| **nous** | parl-ons /paʁlɔ̃/ | parl-âmes /paʁlɑm/ | parl-ions /paʁljɔ̃/ | parl-erons /paʁləʁɔ̃/ | parl-ions /paʁljɔ̃/ | parl-assions /paʁlasjɔ̃/ | parl-erions /paʁləʁjɔ̃/ | parl-ons /paʁlɔ̃/ |
| **vous** | parl-ez /paʁle/ | parl-âtes /paʁlɑt/ | parl-iez /paʁlje/ | parl-erez /paʁləʁe/ | parl-iez /paʁlje/ | parl-assiez /paʁlasje/ | parl-eriez /paʁləʁje/ | parl-ez /paʁle/ |
| **ils** | parl-ent /paʁl/ | parl-èrent /paʁlɛːʁ/ | parl-aient /paʁlɛ/ | parl-eront /paʁləʁɔ̃/ | parl-ent /paʁl/ | parl-assent /paʁlas/ | parl-eraient /paʁləʁɛ/ | |

# Inflections: Nouns

**English** A typical noun has 2 forms: *cat, cats*

- 1 feature: *Number* with 2 values: *Singular, Plural*.

**Swedish** typically 8 forms: *stol, stolen, stolar, stolarna, stols, stolens, stolars, stolarnas*
*äpple, äpplet, äpplen, äpplena, äpples, äpplets, äpplens, äpplenas*

- *Number* with 2 values: *Singular, Plural*.
- *Species* with 2 values: *Indefinite, Definite*
- *Case* with 2 values: *Nominative, Genitive*
- *Gender* with 3 values: *Utrum, Neutrum, Masculine*

**Finnish** 2253 forms of *kauppa* (shop) listed at

`http://www.ling.helsinki.fi/~fkarlsso/genkau2.html`

# Inflections: Nouns

Some forms of the Hungarian noun `ablak` (window):

| | |
|---|---|
| ablaka | its window |
| ablakában | in its window |
| ablakából | from its window |
| ablakai | its windows |
| ablakaik | their windows |
| ablakaikkal | with their windows |
| ablakainak | for their windows |
| ablakán | on its window |
| ablakát | its window (accusativus) |
| ablakba | into the window |
| ablakhoz | towards the window |
| ablakkal | with the window |
| ablakok | windows |
| ablakokat | windows (accusativus) |
| ablakokba | into the windows |
| ablakokkal | with the windows |
| ablakokon | on the windows |
| ablakon | on the window |
| ablakot | window (accusativus) |

# Derivations

By *derivation*, new words can be formed (often from a word of another class)

- *black* $\Rightarrow$ *blackness*
- *affect* $\Rightarrow$ *affection* $\Rightarrow$ *affectionate*
- *compute* $\Rightarrow$ *computation* $\Rightarrow$ *computational*
- (but not *eat* $\Rightarrow$ *eatation**)

Some Swedish examples:

- *motiv* $\Rightarrow$ *motivera* $\Rightarrow$ *motivering*
- *god* $\Rightarrow$ *godhet*
- *äta* $\Rightarrow$ *ätbar*

# Compounds

Swedish can form long compound words, e.g. *hårdvarukompatibilitetsproblem*

Some variations:

- Vowel changes: *hårdvar u kompatibilitetsproblem*
- Vowel drop: *läkarmottagning*
- Extra *s*: *hårdvarukompatibilitet s problem*
- Nothing at all: *tidrapport*

Some compound words have been *lexicalized* (e.g. *football*).

The longest lexicalized Swedish compound word (according to *SAOL*) is *realisationsvinstbeskattning*. (28 letters)

# Morphological analysis and generation

**Morphological analysis** Word form ⇒ lemma + features

- *cats* = `NOUN:cat + NUMBER:plural`
- *stolarnas* = `NOUN:stol + GENDER:utrum + NUMBER:plural + SPECIES:definite + CASE:genitive`

**Morphological generation** Lemma + features ⇒ word form

# SUC (Stockholm-Umea Corpus) tag set

| Feature | Value | Legend | Parts-of-speech where feature applies |
|---|---|---|---|
| Gender | UTR | Uter (common) | DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO) |
| | NEU | Neuter | |
| | MAS | Masculine | |
| Number | SIN | Singular | DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO) |
| | PLU | Plural | |
| Definiteness | IND | Indefinite | DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO) |
| | DEF | Definite | |
| Case | NOM | Nominative | JJ, NN, PC, PM, (RG, RO) |
| | GEN | Genitive | |
| Tense | PRS | Present | VB |
| | PRT | Preterite | |
| | SUP | Supinum | |
| | INF | Infinite | |
| Voice | AKT | Active | |
| | SFO | S-form (passive or deponential) | |
| Mood | KON | Subjunctive (Sw. konjunktiv) | |
| Participle form | PRS | Present | PC |
| | PRF | Perfect | |
| Degree | POS | Positive | (AB), JJ |
| | KOM | Comparative | |
| | SUV | Superlative | |
| Pronoun form | SUB | Subject form | PN |
| | OBJ | Object form | |
| | SMS | Compound (Sw. sammansättningsform) | All parts-of-speech |

# What's the point?

Spell checking (what is a correct word?)

Grammar checking (of *agreement*)

- *The cat s were hungry.* (number agreement)
- *Den svart a svan en* (species agreement)
- *De n svarta svan en* (gender agreement)
- *De n svarta svan en* (number agreement)

Information retrieval By *splitting compound words* and *removing suffixes*, more relevant documents can be found.

- *hårdvarukompatibilitetsproblemen* $\Rightarrow$ *hårdvaru-kompatibilitets-problem-en*

Translation, text generation

# Lemmatization

*Lemmatization* is the process of rewriting words into their lemmata.

*The boys are taller than the girls.* $\rightarrow$
*The boy be tall than the girl.*

Often useful in machine learning contexts where we want to reduce the number of dimensions.

# Lemmatization

For English, a lemmatizer can simply be a look-up table.

```
...
ask      ask
asked    ask
asking   ask
asks     ask
...
```

However, in many languages this solution is not sufficient.