DD2417 Language Engineering
1a: Intro

Johan Boye, KTH

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!

Language engineering applications

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

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...

Underspecification and world knowledge

My car broke down, so I took it to the garage. The guy said that the carburetor was bust, and needed to be replaced.

“Uh-oh”, I thought, “there goes my new computer”.

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)

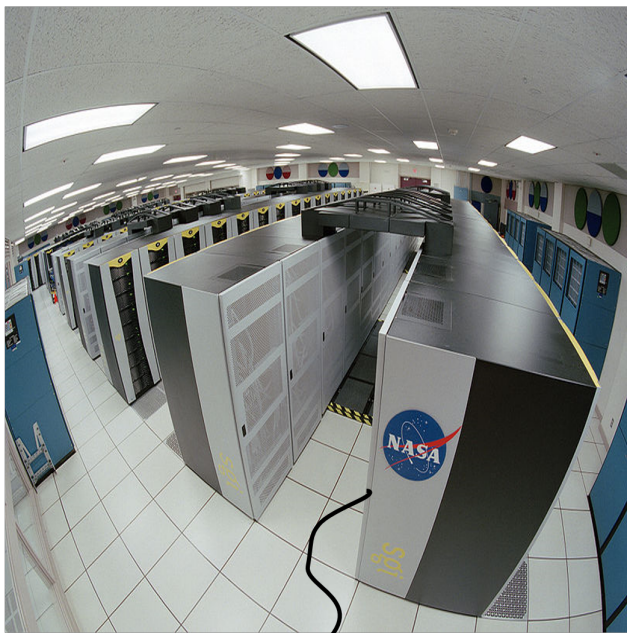
Statistical / “traditional” machine-learning approaches

- 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

The Turing test (1950)



Eliza (Weizenbaum, 1966)

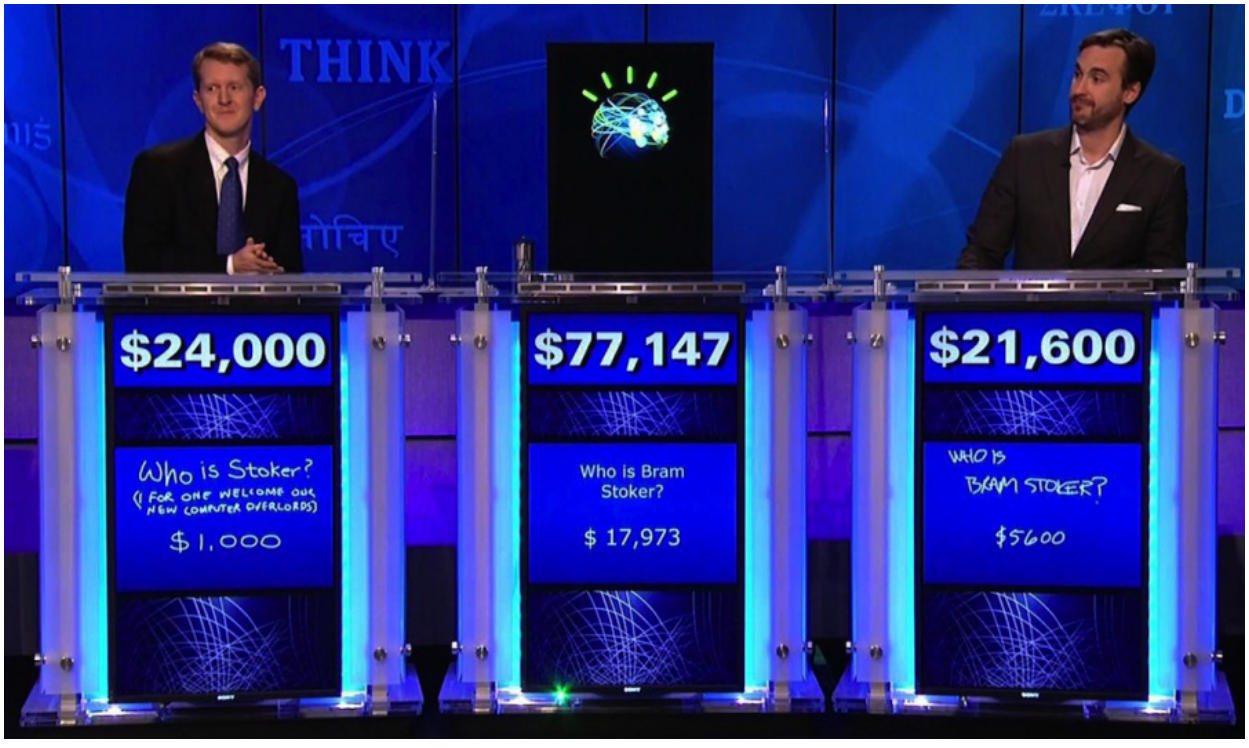
U: I am lonely
E: Are you lonely often?
U: Yes, ever since I was a child
E: What happened when you were a child?
U: Nobody wanted to play with me
E: Why do you say nobody wanted to play with you?

Google Translate (2006,2016)

2006: statistical machine translation.

2016: neural machine translation.

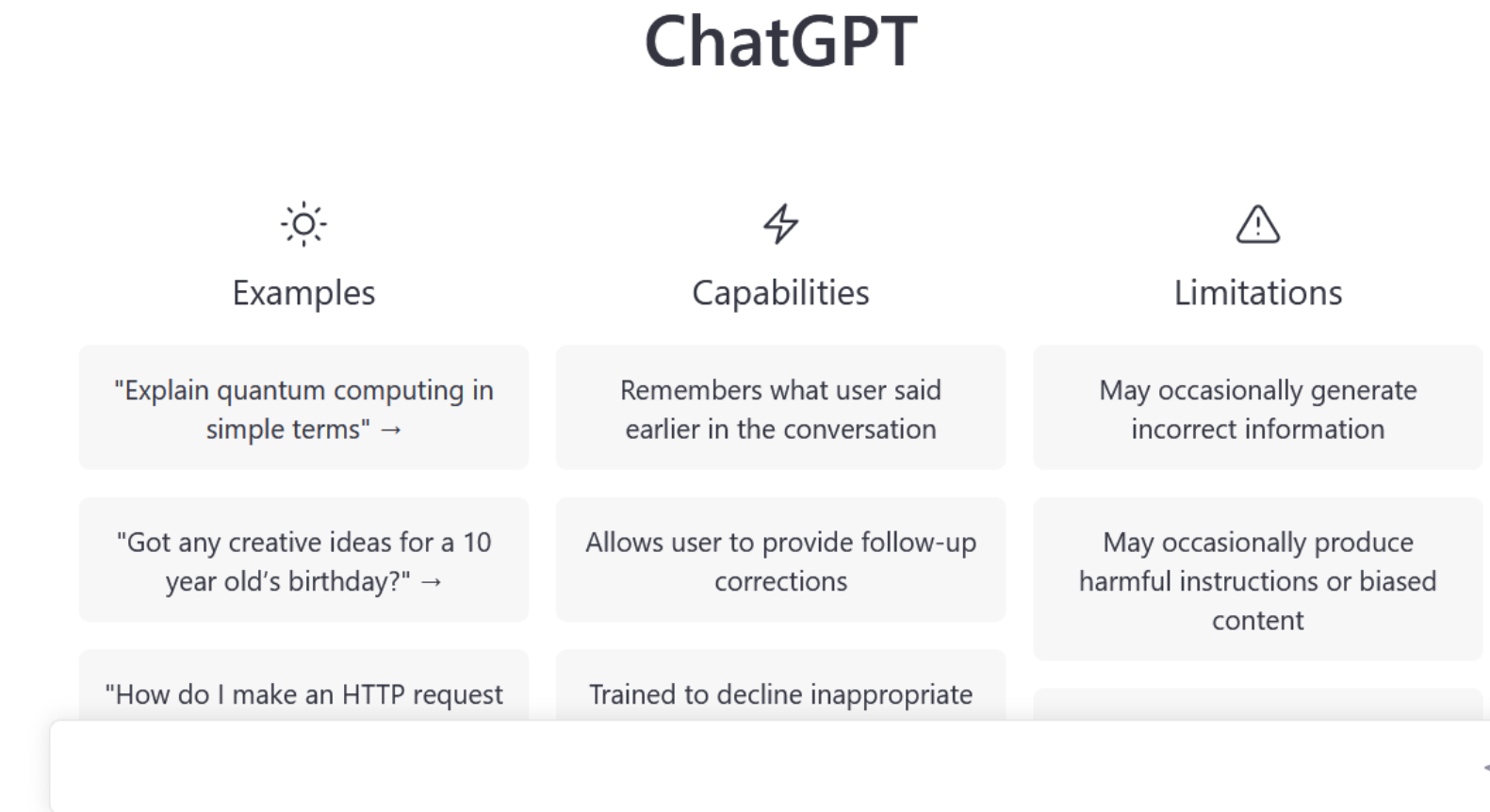
Watson (2011)



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

https://vimeo.com/222234104

ChatGPT (2022)



Large Language Model risks

- Author attribution
- Who's responsible for LLM output?
- Fake news and disinformation
- Death of the middle class
- Superintelligent AI will kill us all?

Course topics

- Grammatical approaches
  - Lecture 1-2, assignment 1
- Statistical approaches
  - Lecture 3-4, assignment 2
- Word as vectors
  - Lecture 5-6, assignment 3
- Neural approaches
  - Lecture 7-9: basics, RNNs, transformers
  - Assignment 4

Teaching

- 9 lectures
  - the two first in-class, the rest digital
- 3 problem-solving seminars
- virtual office hours (on Zoom)
- discussion forum on Canvas

How to get help during office hours

1. Start a Zoom meeting.
2. Place yourself in the queue.
   - Visit https://queue.csc.kth.se/Queue/Sprakt
   - (The queue is locked right now, but will be open during office hours).
   - In the "Location" box, paste your Zoom link.
   - Press the "Join queue" button.
3. The TA will help students in order.

Discussion forum

- Use the discussion forum to ask (and answer) questions
- ... but **don't post code**
- Keep the discussion on a conceptual level.

Examination

4 individual computer assignments

- Each of these has a mandatory part and an optional part
- Grade E: Do all mandatory parts
- Grade D: E + do 1 optional part
- ...
- Grade A: E + do all 4 optional parts

3 quizzes

1 mini-project

- Grade A-F
- Work in pairs

Honor code

The individual assignments are individual:

- Don’t look at somebody else’s code
- Don’t show your code to anyone
- Don’t post your code online

Grading

Final grade will be a combination of the computer assignments grade and the project grade.

Computer assignments	Project				
	A	**B**	**C**	**D**	**E**
A	A	A	B	B	C
B	B	B	B	C	C
C	B	C	C	C	D
D	C	C	D	D	D
E	C	D	D	E	E

Textbook

D. Jurafsky and J. Martin, *Speech and language processing*, 3rd ed. (draft)

Available for free at https://web.stanford.edu/ jurafsky/slp3/

DD2417 Language Engineering
1b: Language

Johan Boye, KTH

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!

2417 Language Engineering
1c: Words

Johan Boye, KTH

Levels of linguistic analysis

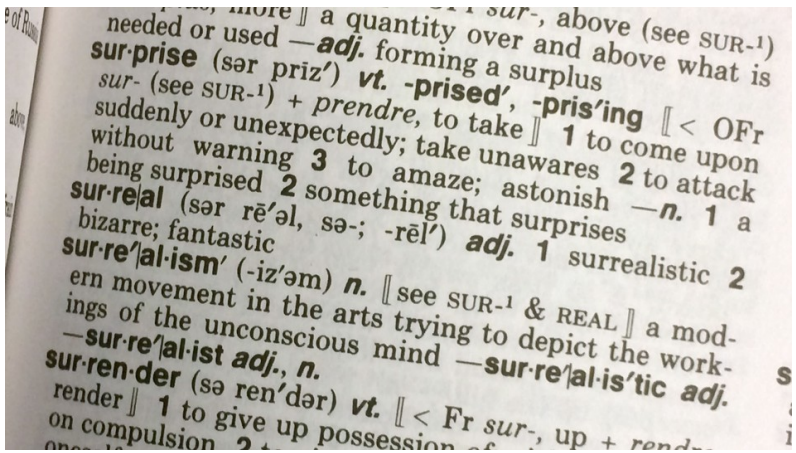
Words	Morphology, Phonology
Sentences	Syntax
Meaning	Semantics
Language use	Pragmatics

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).

We can also use subword tokenization, to reduce the size of the vocabulary, e.g.:

unfriendly $\rightarrow$ un-friend-ly

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.*”

2417 Language Engineering
1d: The structure of words

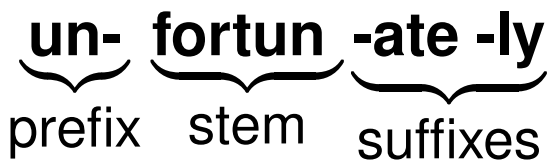
Johan Boye, KTH

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*

Quiz

Determine the morphemes of the words:

- *arrival*
- *employee*

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: 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)

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.