Chapter NLP:III

III. Words

- Word-level Phenomena
- Text Representation
- Text Preprocessing
- Morphological Analysis
- Word Classes

Word Classes

Word class

- In grammar, a word class is a set of words which display the same formal properties, especially in terms of their inflection and distribution.
- Alternative terminology: part of speech, grammatical category, lexical category, syntactic category (roughly synonymous; we won't go into detail).
- The two major types of word classes are:
  - lexical (open/form) classes: nouns, verbs, adjectives, adverbs
  - function (or closed or structure) classes: determiners, particles, prepositions, and others

"When linguists began to look closely at English grammatical structure in the 1940s and 1950s, they encountered so many problems of identification and definition that the term part of speech soon fell out of favor, word class being introduced instead. Word classes are equivalent to parts of speech, but defined according to strict linguistic criteria."

(Crystal (2003) The Cambridge Encyclopedia of the English Language. CUP); definitions below from same source

Word Classes

Open vs. Closed Classes

- Open (lexical words): Theoretically, infinitely many members per class.

Word class	Definition	Comments
Noun	A noun is a word used for naming some person or thing. Examples: man, house, Paris, height	The notional definition is difficult to work with; some grammars add a separate reference to places, but even that excludes many nouns which could not easily be described as 'persons, places, and things', such as abstract qualities (beauty) and actions (a thump). No reference is made to morphology or syntax.
Adjective	An adjective is a word used to qualify a noun. . . to restrict the application of a noun by adding something to its meaning. Examples: fine, brave, three, the	The definition is too broad and vague, as it allows a wide range of elements (e.g, the, my, all) which have very different grammatical properties, and even nouns in certain types of construction (e.g. her brother the butcher) do not seem to be excluded. No reference is made to morphology or syntax
Verb	A verb is a word used for saying something about some person or thing. Examples: make, know, buy, sleep	On this definition, there is little difference between a verband an adjective (above). Some grammars prefer to talk about 'doing words' or 'action words', but this seems to exclude the many state verbs, such as know, remember, and be. No reference is made to morphology or syntax.
Adverb	An adverb is a word used to qualify any part of speech except a noun or pronoun. Examples: today, often, slowly, very	This is an advance on the more usual definition, in which adverbs are said to qualify (or 'modify') verbs – which is inadequate for such words as very and however. Even so, the definition leaks, as it hardly applies to interjections, and examples such as the very man and slovenly me have to be thought about. Nothing is really said about morphology or syntax

Word Classes

Open vs. Closed Classes

- Closed (function words): Number of members is fixed in principle.

  As language evolves, changes may rarely also happen in closed classes.

Word class	Definition	Comments
Pronoun	A pronoun is a word used instead of a noun or noun-equivalent (i.e. a word which is acting as a noun). Examples: this, who, mine	The definition is almost there, but it has to be altered in one basic respect: pronouns are used instead of noun phrases, not just nouns. He refers to the whole of the phrase the big lion, not just the word lion (we cannot say *the big he). Nothing is said about morphology or syntax
Preposition	A preposition is a word placed before a noun or noun-equivalent to show in what relation the person or thing stands to something else. Examples: on, to, about, beyond	This is a good start, as it gives a clear syntactic criterion. The definition needs tightening up, though, as prepositions really go before noun phrases, rather than just nouns, and may also be used in other parts of the sentence. As with nouns above, more than just persons and things are involved.
Conjunction	A conjunction is a word used to join words or phrases together, or one clause to another clause, Examples: and, before, as well as	This captures the essential point about conjunctions, but it also needs some tightening up, as prepositions might also be said to have a joining function (the man in the garden). A lot depends on exactly what is being joined
Interjection	An interjection is a word or sound thrown into a sentence to express some feeling of the mind. Examples: Oh!, Bravo!, Fie!	This is vaguer than it need be, [. . .] the essential point [being] that interjections do not enter into the construction of sentences. Despite the emotional function of these words, they still need to be considered as part of sentence classification

Word Classes

Ambiguities

- About 90% of all known wordforms belong to only one word class.
- The others are ambiguous

  The back door → adjective, JJ
  On my back → noun, NN
  Win the voters back → adverb, RB
  Said to back the bill → verb, VB

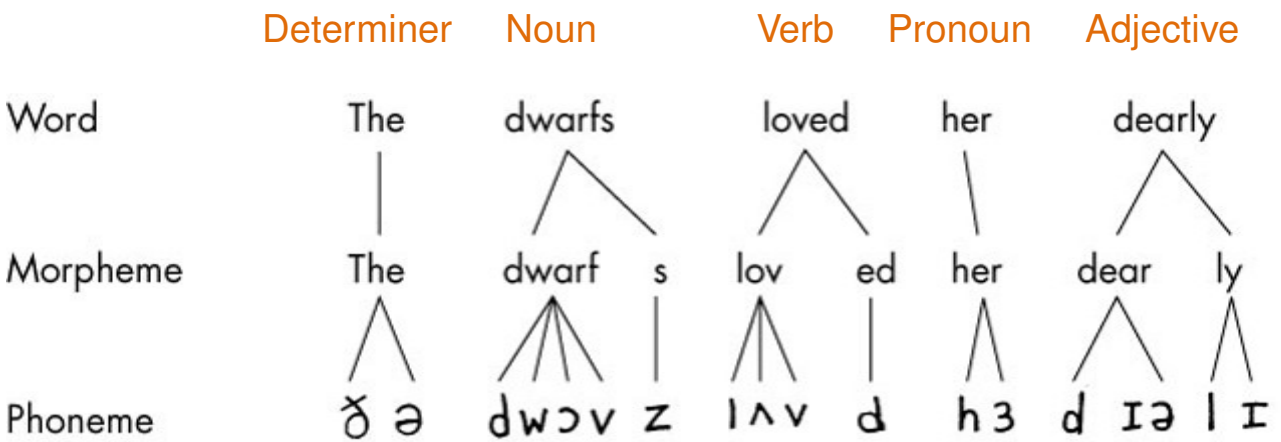
- Analysis of syntactic context helps disambiguate.

"There is no single correct way of analyzing words into word classes. . . Grammarians disagree about the boundaries between the word classes, and it is not always clear whether to lump subcategories together or to split them. For example, in some grammars pronouns are classed as nouns, whereas in other frameworks they are treated as a separate word class." Aarts, Chalker, Weiner (2014) The Oxford Dictionary of English Grammar. OUP

Word Classes

Part-of-Speech Tagging

- The process of assigning a word class/part of speech from a predefined set to a word.



Word Classes

Part-of-Speech Tagging

Given a token sequence of text, markup each token with its part of speech (POS).

Common English (Western) 9 parts of speech:

- Noun names of abstract or concrete entities: persons, places, things, ideas, qualities
- Pronoun substitutes for nouns
- Verb actions, occurrences, or states of being
- Adjective modifiers of a noun or pronoun
- Adverb modifiers of verbs, adverbs, or adjectives
- Preposition words expressing relations in a phrase or sentence
- Conjunction connects words, phrases, or clauses
- Interjection expressions of feelings and emotions
- Determiner including articles (markers of definiteness or indefiniteness), demonstratives (pointing “this”, “that”), possessive determiners (“my”, “her”), quantifiers (“all”, “few”)

For practical purposes, these broad classes are insufficient. Typically some 30 to 150 parts of speech are distinguished.

Word Classes

Part-of-Speech Tagging

Given a token sequence of text, markup each token with its part of speech (POS).

For practical purposes, typically 30 to 160 parts of speech are distinguished; the concrete set of parts of speech used is referred to as tag set[1]

- Penn Treebank tagset — 36 tags
- CLAWS tagsets — CLAWS1: 132, CLAWS2: 166, C5: 60, C6: 160, C8, ...
- Universal POS tags — 17 tags

[1] also spelled as "tagset"

Word Classes

Part-of-Speech Tagging

Given a token sequence of text, markup each token with its part of speech (POS).

Universal tagset: core part-of-speech categories; universally applicable

Open class

ADJ: adjective
ADV: adverb
INTJ: interjection
NOUN: noun
PROPN: proper noun
VERB: verb

Closed class

ADP: adposition
AUX: auxiliary
CCONJ: coordinating conjunction
DET: determiner
NUM: numeral
PART: particle
PRON: pronoun
SCONJ: subordinating conjunction

Other

PUNCT: punctuation
SYM: symbol
X: other

Word Classes

Part-of-Speech Tagging

Given a token sequence of text, markup each token with its part of speech (POS).

Penn Treebank tagset

CC Coordinating conjunction
CD Cardinal number
DT Determiner
EX Existential there
FW Foreign word
IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS List item marker
MD Modal
NN Noun, singular or mass
NNS Noun, plural
NNP Proper noun, singular
NNPS Proper noun, plural
PDT Predeterminer
POS Possessive ending
PRP Personal pronoun
PRP$ Possessive pronoun
RB Adverb
RBR Adverb, comparative
RBS Adverb, superlative
RP Particle
SYM Symbol
TO to
UH Interjection
VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non-3rd person singular present
VBZ Verb, 3rd person singular present
WDT Wh-determiner
WP Wh-pronoun
WP$ Possessive wh-pronoun
WRB Wh-adverb

Word Classes

Part-of-Speech Tagging

STTS (Stuttgart-Tuebingen-Tagset)

pos-tag	Beschreibung	Beispiel(e)
ADJA	attributives Adjektiv	*der schlaue/ADJA Mitarbeiter*
ADJD	adverbiales ODER	*er spricht schnell/ADJD*
	prädikatives Adjektiv	*Sein Sprechen ist schnell/ADJD*
ADV	Adverb	*Bald/ADV schon/ADV kommt sie wohl/ADV*
APPR	Präposition; Zirkumposition links	*nach/APPR Berlin; ohne/APPR Hund*
APPRART	Präposition mit Artikel	*zum/APPRART Streichen; zur/APPRART Sache*
APPO	Postposition	*ihm zuliebe/APPO; der Sache wegen/APPO*
APZR	Zirkumposition rechts	*von mir aus/APZR*
ART	bestimmter ODER	*Der/ART Mann schenkt die/ART Rose*
	unbestimmter Artikel	*einer/ART unerwarteten Frau*
CARD	Kardinalzahl	*zwei/CARD Männer im Jahre 1994/CARD*
FM	Fremdsprachliches Material	*Er sagte:" Hasta/FM luego/FM, amigos/FM ."*
ITJ	Interjektion	*Mhm/ITJ, ach/ITJ, tja/ITJ, dann halt nicht.*
KOUI	unterordnende Konjunktion mit (*zu*-)Infinitiv	*Sie kommt, um/KOUI zu arbeiten*
		Anstatt/KOUI anzufangen, geht sie wieder
KOUS	unterordnende Konjunktion	*Emma wartet, weil/ob/solange/dass/KOUS sie stiehlt*
KON	nebenordnende Konjunktion und, oder, aber	*Sie und/oder/KON Emma kommen und/KON streichen*
KOKOM	Vergleichskonjunktion als, wie	*blauer als/KOKOM er; blau wie/KOKOM er*
NN	normales Nomen	*am Tage/NN dem Mann/NN den Schlaf/NN*
NE	Eigennamen	*die Emma/NE dem Hans/NE sein HSV/NE*

https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger

Word Classes

Part-of-Speech Tagging

Given a token sequence of text, markup each token with its part of speech (POS).

CLAWS C8 tagset

see http://ucrel.lancs.ac.uk/claws8tags.pdf for the **170** tags

Word Classes

Part-of-Speech Tagging

Original text:

```
A relevant document will describe marketing strategies carried out by
U.S. companies for their agricultural chemicals, report predictions for
market share of such chemicals, or report market statistics for
agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer,
predicted sales, market share, stimulate demand, price cut, volume of
sales.
```

Brill tagger:

```
A/DT relevant/JJ document/NN will/MD describe/VB marketing/NN
strategies/NNS carried/VBD out/IN by/IN U.S./NNP companies/NNS for/IN
their/PRP$ agricultural/JJ chemicals/NNS ,/, report/NN predictions/NNS
for/IN market/NN share/NN of/IN such/JJ chemicals/NNS ,/, or/CC report/NN
market/NN statistics/NNS for/IN agrochemicals/NNS ,/, pesticide/NN ,/,
herbicide/NN ,/, fungicide/NN ,/, insecticide/NN ,/, fertilizer/NN ,/,
predicted/VBN sales/NNS ,/, market/NN share/NN ,/, stimulate/VB demand/NN
,/, price/NN cut/NN ,/, volume/NN of/IN sales/NNS ./.
```

CC	coordinating conjunction	NN	singular or mass noun	VBD	verb, past tense
DT	singular determiner/quantifier	NNP	proper noun, singular	VBN	verb, past participle
IN	preposition	NNS	plural noun	,	comma
JJ	adjective	PRP$	possessive pronoun	.	dot
MD	modal auxiliary	VB	verb, base form		other tags

Word Classes

Part-of-Speech Tagging

apple (single noun, NN), apples (plural noun, NNS), Apple (proper noun, NNP),
sigh (verb base form, VB), sighed (verb past tense *or* past participle, VBD or VBN),
the (determiner, DT), it (personal pronoun, PRP), WHATZ (???), ...

Part-of-speech tagged data provides valuable information about

- a word and its possible neighbors
- the correct pronunciation (speech synthesis)
  - OBject vs. obJECT, CONtent vs. conTENT
- its intended sense (word sense disambiguation)
- the applied morphemes (lemmatization)
- the meaning of a sentence (shallow parsing)

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1992]

Principle: “error-driven transformation-based tagging”

1. Assign each token its most likely part of speech tag. Stemming rules are applied to match inflected tokens with word stems stored in a dictionary.
2. Apply a list of transformation rules to correct tagging errors.
3. Repeat Step 2 until no rules can be applied, anymore, or after a pre-specified number of repetitions.

Concepts:

- Initial tag probabilities are trained on a large pre-tagged corpus.
- Rules are learned from errors made on a pre-tagged corpus, and applied in the order listed.
- Rules are defined as follows: `T1 T2 <Premise>`

Semantics:

For each token currently tagged with `T1` which fulfills the `<Premise>`, replace `T1` with `T2`.

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1992]

Premises:

`context x`	A word in `context` is tagged `x`.
`property`	The word has a certain `property`.
`context property`	A word in `context` has a certain `property`.
`context`	One or any of $i \in [1, 3]$ preceding or following word(s).
`property TRUE|FALSE`	Capitalized word.

Example rules:

```
TO     IN     next-tag AT // I like to go.  vs.  I go to the cinema.
VBN    VBD    prev-word-is-cap TRUE
VBD    VBN    prev-1-or-2-or-3-tag HVD
VB     NN     prev-1-or-2-tag AT
NN     VB     prev-tag TO
TO     IN     next-word-is-cap TRUE
NN     VB     prev-tag MD
```

Rules are learned starting with the initial tagging on a training dataset by instantiating rules from the above templates, keeping those that minimize tagging errors the most in each iteration, until some termination criterion is reached.

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1994]

Problem: The tagger cannot tag words not occurring in the training data.

An unknown word tagger can be trained based on the same principles but with different premises as templates for rules. `T1` may be `UNK` for unknown.

Premises:

`affix x constraint`	Token fulfills `constraint` regarding affix of at most 4 chars.
`context word`	A word appears in `context`.
`char x`	Character `x` occurs in word.
`constraint`	When deleting or adding affix `x`, word found in dictionary. Else, affix `x` occurs in token.

Example rules:

```
NN     NNS    suffix -s occurs
NN     CD     char .
NN     JJ     char -
NN     VBN    suffix -ed occurs
NN     VBG    suffix -in occurs
UNK    ADJ    suffix -ly addition
UNK    RB     suffix -ly occurs
```

Remarks:

- Large corpora for part of speech tagging have been painstakingly manually annotated, starting with the 1 million word Brown corpus in the 1960s, later superseded by the 100 million word British National Corpus, and others.
- Tag sets: Brown (87 tags), Penn TreeBank II (41 tags), British National Corpus (61 tags), British National Corpus Sampler (146 tags).
- Assigning the most probable tag to each known word and proper noun to all unknown words already yields 90% accuracy. [Charniak 1997]
- The state of the art in part of speech tagging can be reviewed at aclweb.org. Most taggers reported are based on statistical sequence models rather than rules. However, many taggers proposed are not included, including the Brill tagger.
- Nevertheless, the Brill tagger frequently serves as baseline for comparison, and as a last step in tagging pipelines.