Chapter NLP:IV

IV. Words

Word Classes
Definition

A word class is a set of lexical items with similar formal (grammatical) properties.

- ❑ Also called part of speech, grammatical category, lexical category, or syntactic category. roughly synonymous
- ❑ Common properties are morphology and semantic or syntactic behavior.
- ❑ Often serve an encoding of the formal properties of words for programs.

	Determiner	Noun	Verb	Pronoun	Adjective
Word	The	dwarfs	loved	her	dearly
Morpheme	The	dwarf s	lov ed	her	dear ly
Phoneme	ð ə	dwɔv z	lʌv d	h3	d ɪə l ɪ
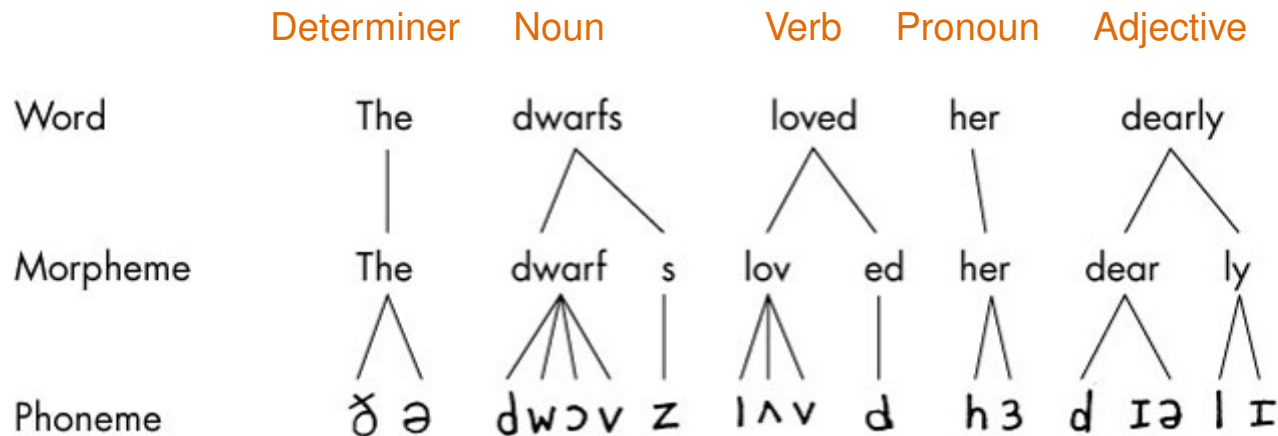
Word Classes

Traditional grammar

The traditional English grammar lists 9 word classes split into two groups: function or content classes.

1. Content classes: or open/form/lexical

   ❑ Accept new members, can hold infinitely many items per class.

   ❑ Nouns, verbs, adjectives, adverbs.

2. Function classes: or closed/structure

   ❑ Number of members is fixed.

   ❑ Prepositions, pronouns, determiners, conjunctions and interjections.

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

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN					
UD					

Word Classes

Traditional grammar: Example

Word Class: Noun
Definition: A noun is a word used for naming some person or thing. [Cambridge]
Examples: Paris, man, house, height.

❑ The definition is incomplete: What about places? What about abstract qualities (beauty) and actions (a thump)?

❑ No reference is made to morphology or syntax.

❑ The class does not differentiate (grammatical) properties: Plural nouns and proper nouns are all just **noun**.

Remarks:

- ❑ 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 vague and allows many elements with different grammatical properties (`the, my, all`), and even nouns in certain constructions (`her brother the butcher`).
  - – No reference is made to morphology or syntax.

- ❑ 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 verb and an adjective. Some grammars prefer to talk about 'doing words' or 'action words', but this seems to exclude the many state verbs, such as `know, remember, be`.
  - – No reference is made to morphology or syntax.

- ❑ An **adverb** is a word used to qualify any part of speech except a noun or pronoun. Examples: `today, often, slowly, very`.
  - – Adverbs are often said to qualify (or 'modify') verbs – which is inadequate for such words as `very` and `however`.
  - – This definition hardly applies to interjections or examples like `the very man` and `slovenly me`.
  - – No reference is made to morphology or syntax.

Remarks:

❑ 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.`

    – 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`, since we cannot say `the big he`.

    – No reference is made to morphology or syntax.

❑ 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 gives a clear syntactic criterion. However, prepositions go before noun phrases, rather than nouns, and may also be used in other parts of the sentence. As with nouns, more than just persons and things are involved.

❑ A **conjunction** is a word used to join words or phrases together, or one clause to another clause, Examples: `and, before, as well as.`

    – Prepositions also have a joining function (`the man in the garden`).

    – Conjunction Junction, what's your function?

❑ An **interjection** is a word or sound thrown into a sentence to express some feeling of the mind. Examples: `Oh!, Bravo!, Fie!.`

    – 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
Tagsets

The classes of the traditional grammar are not suited for language processing.

❑ Several advanced class schemes (called tagsets) exist.

❑ They distinguish between 17 up to 100+ word classes.

- Penn Treebank tagset   36 tags
- Universal POS tags   17 tags
- CLAWS tagsets   CLAWS1: 132, CLAWS2: 166, C5: 60, C6: 160, C8, …
- Brown Corpus   87 tags
- Penn Treebank II   41 tags
- British National Corpus   61 tags
- BNC Sampler   146 tags

❑ Corpora for part of speech tagging are manually annotated, for example:

- The 1 million word Brown corpus in the 1960s.
- The 100 million word British National Corpus.

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties.

NN	Noun in singular
NNS	Noun in plural
NNP	Proper noun
NNPS	Proper noun in plural

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties for all classes.

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
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
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
WRB	Wh-adverb

TO	to (Preposition)
RP	Particle
POS	Possessive ending
MD	Modal
PRP	Personal pronoun
PRP$	Possessive pronoun
WP	Wh-pronoun
WP$	Possessive wh-pronoun
DT	Determiner
PDT	Predeterminer
WDT	Wh-determiner
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
UH	Interjection

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties for all classes. Add classes for everything else we find while annotating text.

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
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
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
WRB	Wh-adverb

TO	to
RP	Particle
POS	Possessive ending
MD	Modal
PRP	Personal pronoun
PRP$	Possessive pronoun
WP	Wh-pronoun
WP$	Possessive wh-pronoun
DT	Determiner
PDT	Predeterminer
WDT	Wh-determiner
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
UH	Interjection
CD	Cardinal number
EX	Existential there
FW	Foreign word
LS	List item marker
SYM	Symbol

Word Classes

Penn Treebank tagset [upenn]

The Penn tagset covers more grammatical properties and is still frequently used.

The Penn tagset is also English-centric:

- ❑ What does it mean and why is this a problem?

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD					

Word Classes

Penn Treebank tagset [upenn]

The Penn tagset covers more grammatical properties and is still frequently used.

The Penn tagset is also English-centric:

- ❑ With this strategy, each language needs it's own tagset.
- ❑ A different tagger needs to be developed for each language.
  Very difficult for low-resource languages, see electronic colonialism.
- ❑ It violates Chomsky's theory of universal grammar, which is very beloved by computer scientists.

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD					

Word Classes

Universal Dependencies tagset [UD]

Idea: Seperate form from function: Use only few, universal POS tags and many lexical and grammatical properties which can be freely assigned to any tag.

Part-of-speech Tag	
NOUN	Noun
PROPN	Proper noun
VERB	Verb
AUX	Auxiliary
ADJ	Adjective
ADV	Adverb
ADP	Adposition (Preposition)
PART	Particle
PRON	Pronoun
DET	Determiner
CCONJ	Coordinating conjunction
SCONJ	Subordinating conjunction
INTJ	Interjection
NUM	Numeral
PUNCT	Punctuation
SYM	Symbol
X	Other

Lexical Features
PronType
NumType
Poss
Reflex
Foreign
Abbr
Typo

Inflectional Features	
Nominal	Verbal
Gender	VerbForm
Animacy	Mood
NounClass	Tense
Number	Aspect
Case	Voice
Definite	Evident
Degree	Polarity
	Person
	Polite
	Clusivity

Which of these do not exist in English?

Word Classes

Universal Dependencies tagset [UD]

Idea: Seperate form from function: Use only few, universal POS tags and many lexical and grammatical properties which can be freely assigned to any tag.

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD	DET; PronType: Art	NOUN; Number: Plur	VERB; Tense: Past	PRON	ADV

Word Classes
Ambiguities

- About 85% of the vocubulary (types) 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

- However, ambiguous types are much more frequent. About half of the tokens are ambiguous.

10 REASONS WHY ENGLISH IS WEIRD

1) The bandage was wound around the wound.
2) The farm was used to produce produce.
3) The dump was so full that it had to refuse more refuse.
4) We must polish the Polish furniture.
5) He could lead if he would get the lead out.
6) The soldier decided to desert his dessert in the desert.
7) Since there is no time like the present, he thought it was time to present the present.
8) A bass was painted on the head of the bass drum.
9) When shot at, the dove dove into the bushes.
10) I did not object to the object.

Remarks:

❑ About the termininology, the Cambridge Encyclopedia of the English Language (CUP) writes:

"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)

❑ "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

❑ The 9 word classes of the traditional grammar are usually attributed to *Dionysius Thrax of Alexandria (100 B.C.)*, who wrote in his *Techne* on Greek about eight parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article.

Word Classes

Part-of-Speech Tagging

A tagger is a program that assigns tags from a tagset to each token in a sequence. Tagging is a sequence labeling problem: the input is a sequence of tokens $w_1, \ldots, w_n$ and a tagset T and the output is a sequence of tags $c_1, \ldots, c_n$, $c_i \in T$ of equal length.

Part-of-speech tags can inform us about syntax and semantics of a sequence:

- the intended sense of a word

  apple (single noun, NN) vs. Apple (proper noun, NNP)

- the applied morphemes (lemmatization)

  sigh (verb base form, VB), sighed (verb past tense *or* past participle, VBD or VBN)

- the meaning of a sentence (shallow parsing)
- the correct pronunciation (speech synthesis)

  - OBject vs. obJECT, CONtent vs. conTENT

Word Classes

Part-of-Speech Tagging: Maximum Likelihood Estimate

Idea: Most types are unambiguous. Most ambiguous types have one very likely tag. So we can do a Maximum Likelihood Estimate (MLE):

❑ Tag each token w_i with the word class c_i it appears in most often:[1]

$$c_i = \operatorname*{argmax}_{c_j \in T} \frac{C(c_j, w_i)}{C(w_i)}$$

❑ Unknown words are often tagged as proper nouns.

❑ This is sometimes called the most frequent class baseline.

❑ This baseline works extremely well; It achieves 92% accuracy on UD over 15 languages [Wu and Dredze 2019].

❑ Humans reach roughly 97% accuracy.

---

[1]$C(w)$ is the count function.

Word Classes

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

Idea: Improve the most frequent class baseline by applying rules for special cases.

The Brill Tagger is an "error-driven transformation-based tagging". It iteratively applies rules `T1 T2 <Premise>`.

1. Initially tag a sequence with MLE.
2. If a token is tagged with `T1` and fulfills the `<Premise>`, replace `T1` with `T2`.
3. Repeatedly iterate the sequence and apply the rules in order until the stopping criterion is reached.

The rules are learned from errors made on a pre-tagged corpus.

- ❑ If a false tag `T1` is encounterd in the pre-tagged corpus, create several rules with the correct tag `T2` and a fulfilled `<Premise>`.
- ❑ Correct a test corpus with each new rule.
- ❑ If a rule does not increase accuracy, discard it.

Word Classes

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

Premises are constructed from templates:

`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
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
```

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

Idea: Create rules to tag unknown tokens. `T1` may be `UNK` for unknown.

Premise templates for unknown tokens:

`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
```

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.

Tagged text with 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** ./.

Remarks:

- ❏ Part-of-speech tagging can be solved with any generic sequence labeling model, like Hidden Markov Models, CRFs, RNNs, and transformer models.

- ❏ The Brill Tagger can be seen as a special case of a sequence classifier, where the premises are the features and the order of the rules are the weights.

- ❏ The state of the art in part of speech tagging can be reviewed at aclweb.org, paperswithcode, or NLPprogress. Most taggers reported are based on statistical sequence models rather than rules. However, many taggers proposed are not included, including the Brill tagger.