Morphology:
Part of speech tagging
PART 1

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

Overview

- What is morphology? The knowledge annotated in morphological resources

- Tagsets and languages in morphological resources: Penn tag set for English, Universal Dependencies tag set, tag sets for other languages (Turkish, Spanish)

- Resources and tasks for morphological analysis: Kipos, PosTWITA, British National Corpus, Universal Dependencies,

What is morphology?

Morphological knowledge refers to both
the internal structure of words
and their classification according to the functions they
play within the sentence.

Morphology

In traditional linguistics, **morphology** is the study of the structure and classification of the words in a particular language (from ancient Greek μορφή = 'form').

Morphology considers:
- the principles of formation of the **structure** of each single word, how morphemes combine into meaningful units like prefixes, suffixes and roots
- how words can be grouped into classes called **parts of speech**.

A **lexeme** (or lemma) is the basic unit of morphological analysis, the basic form (or the set of forms) taken by a single word or group of words.
A lexeme can be an individual word (talk, talks, talked …) or a multiword expression (speak up, speaks up, …).

The way in which a lexeme is used in a sentence is determined by its **grammatical category** also called part of speech.

The grammatical category of a lexeme provides crucial hints for determining the syntactic position and function of the lexeme within the sentence.

Morphological ambiguity

According to a morphological perspective, a **lexeme** is ambiguous when it can belong to different grammatical categories:

English: light > **noun** in ‘*light bulb*’
adjective in ‘*light as a feather*’

Italian: rosa > **noun** in ‘è *fiorita una rosa*’ (a rose has bloomed)
adjective in ‘*un abito rosa*’ (a pink dress)

POS TAGGED CORPORA

- In NLP, a morphologically annotated corpus is technically defined as a *Part of Speech tagged corpus*
- *A* PoS tagged corpus is a collection of sentences where each **token** is associated with a proper **Part of Speech tag**
- PoS tagging is the classification task which consists in providing a morphological analysis of a single token by selecting the proper Part of Speech in a given set of **grammatical categories and features**, called **tag set**

POS TAGGING

- In different theoretical linguistic frameworks, inventories of **labels** that describe the morphology of words have been proposed which usually agree on the most of **grammatical categories and features**
- In computational linguistic, these inventories were transposed in several different **tag sets**, which meaningfully vary in amount and typology of categories and features, also according to the language for which they were created

POS TAGGING AND LANGUAGES

- Different sets of categories are needed for the annotation of morphology of different languages.

  Example: languages with/without articles

- The same category can be realised in very different ways in different languages.

  Example: verbs with and without inflection

POS TAGGING: CATEGORIES

In theoretical grammar frameworks, the most often used grammatical sets of categories include:

- **Noun** fish, book, house, pen, procrastination, language
- **Proper noun** John, France, Barack, Goldsmiths, Python
- **Verb** loves, hates, studies, sleeps, thinks, is, has
- **Adjective** grumpy, sleepy, happy, bashful
- **Adverb** slowly, quickly, now, here, there
- **Pronoun** I, you, he, she, we, us, it, they
- **Preposition** in, on, at, by, around, with, without
- **Conjunction** and, but, or, unless
- **Determiner** the, a, an, some, many, few, 100

POS TAGGING: CATEGORIES

Different scholars proposed in the past different set of PoS categories. For several centuries, the most used and cited classification was that provided by Dionysius Thrax of Alexandria (100 b.C.), which includes:

- **Noun** (including also Proper noun)
- **Verb**
- **Adverb**
- **Pronoun**
- **Preposition**
- **Conjunction**
- **Participle** (including also adjective?)
- **Article** (not considering other kind of determiner?)

POS TAGGING: CATEGORIES

It is important to consider **Proper noun** as a separate category because it allows us to easily identify named entities, such as people or places, that may play an important role in the events described by verbs.

For example, for many natural language understanding tasks, such as question answering, stance detection, or information extraction, it can be useful to know whether a named entity such as *Torino* is a name of a person, place, or university.

POS TAGGING: CATEGORIES

Not for all categories a sharp distinction can be traced with respect to the other ones.

Also a standard distinction between nouns (as generally referred to people, places, things or concepts) and verbs (as generally referred to events or actions) can be criticised.

Can you find examples where the same concept can be expressed by a noun or a verb, or by an adjective or an adverb?

POS TAGGING: CATEGORIES

Can you find examples where the same concept can be expressed by a noun or a verb, or by an adjective or an adverb?

- *Rome* ***fell swiftly*** */ The* ***fall*** *of Rome was* ***swift****.*
- *The enemy* ***completely*** *destroyed the city. /*
  The enemy's destruction of the city was ***complete****.*
- *John* ***loves*** *Mary too much! / John's* ***love*** *for Mary is eccessive!*

POS TAGGING: CATEGORIES

Not for all words it is easy to decide which category they belong to.

E.g.: It's '*about*' a preposition or an adverb?

We talked about trees

At each lesson there are about 30 students

We can **assign a part of speech to a word**:

- in a rigid and **unchanging** way, regardless the use of the word in the sentence
- depending on its **function within the context** of the sentence

POS TAGGING: CATEGORIES

Some types of words do not perfectly fit with a category where they were traditionally collocated.

Example: verbs corresponds to actions or states, but auxiliary or modal verbs (such as *do*, *shall* and *be*) do not correspond to any particular action/ state, serving a purely grammatical function.

To deal with some cases it seems useful to consider **broader and general categories**, in others it seems better to use more **specific** ones.

Example: considering that articles and demonstrative adjectives play a very similar role, we can include both in the determiner category

POS TAGGING: FEATURES

Some grammatical category may be also defined at finer grain by using an inventory of features:

Number

- associated with noun, verb, pronoun, adjective …
- two values: singular and plural (dog/dogs, this/these, bello/belli, he/they)

Gender

- associated with personal pronoun, noun, verb …
- three values: masculine, feminine and neutral (he, she, it, they, Bella/bello/belli/belle)

POS TAGGING: FEATURES

Person

- associated with verb and pronoun
- three values: 1st (the person who is speaking) , 2nd (who is the hearer) and 3rd (the person or thing about whom we are speaking)

Case

- associated with personal and interrogative pronoun
- three values:

  Nominative with the function of subject (I, we, you, he, she, it, they, who).

  Genitive with the function of possessor (my/mine, our/ours, his, her/hers, its, their/theirs, whose)

  Objective with the function of object (me, us, you, him, her, them, whom)

POS TAGGING: FEATURES

Degree

- associated with adjective and adverb

- three terms:
    - **Positive** which expresses a quality (big, fast, beautiful)
    - **Comparative** which expresses greater intensity of a quality in one of two items (bigger, faster, more beautiful)
    - **Superlative** which expresses greatest intensity of a quality in one of three or more items (biggest, fastest, most beautiful)

POS TAGGING: FEATURES

Definiteness
- associated with article and indefinite adjective

- two values:
  - **Definite** which indicates a referent identifiable by speaker/hearer (the, il, la, gli)
  - **Indefinite** which indicates a referent not precisely identifiable by speaker/hearer (a/an, some, un, una, alcuni)

POS TAGGING: FEATURES

Tense
- associated with verb

- three basic values (with a great variation in more or less flessive languages):

    Present which represents an action in the present moment (John works hard)
    Past which represents an action before the present moment (it rained)
    Future which represents an action after the present moment (it will rain)

POS TAGGING: FEATURES

Aspect
- associated with verb

- two basic values:
  - **Perfective** which describes an action as whole and complete (yesterday I met my friend)
  - **Imperfective** which describes an action as ongoing and incomplete (I am working on this problem)

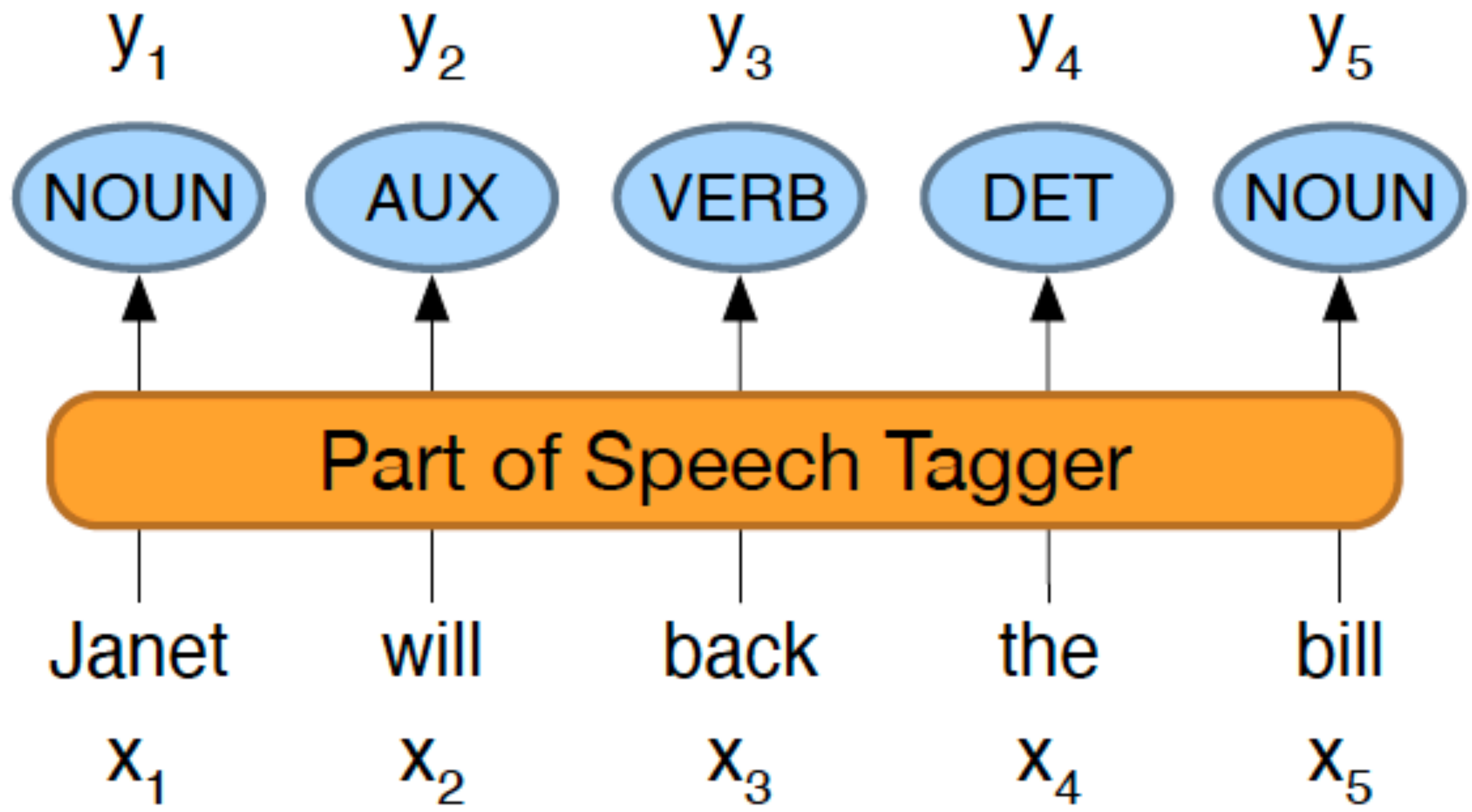
POS TAGGING: FEATURES

Mood
- associated with verb for showing the speaker’s attitude towards what he/she is talking about

- some values: indicative, interrogative, imperative, injunctive, subjunctive, potential, optative, gerund, …

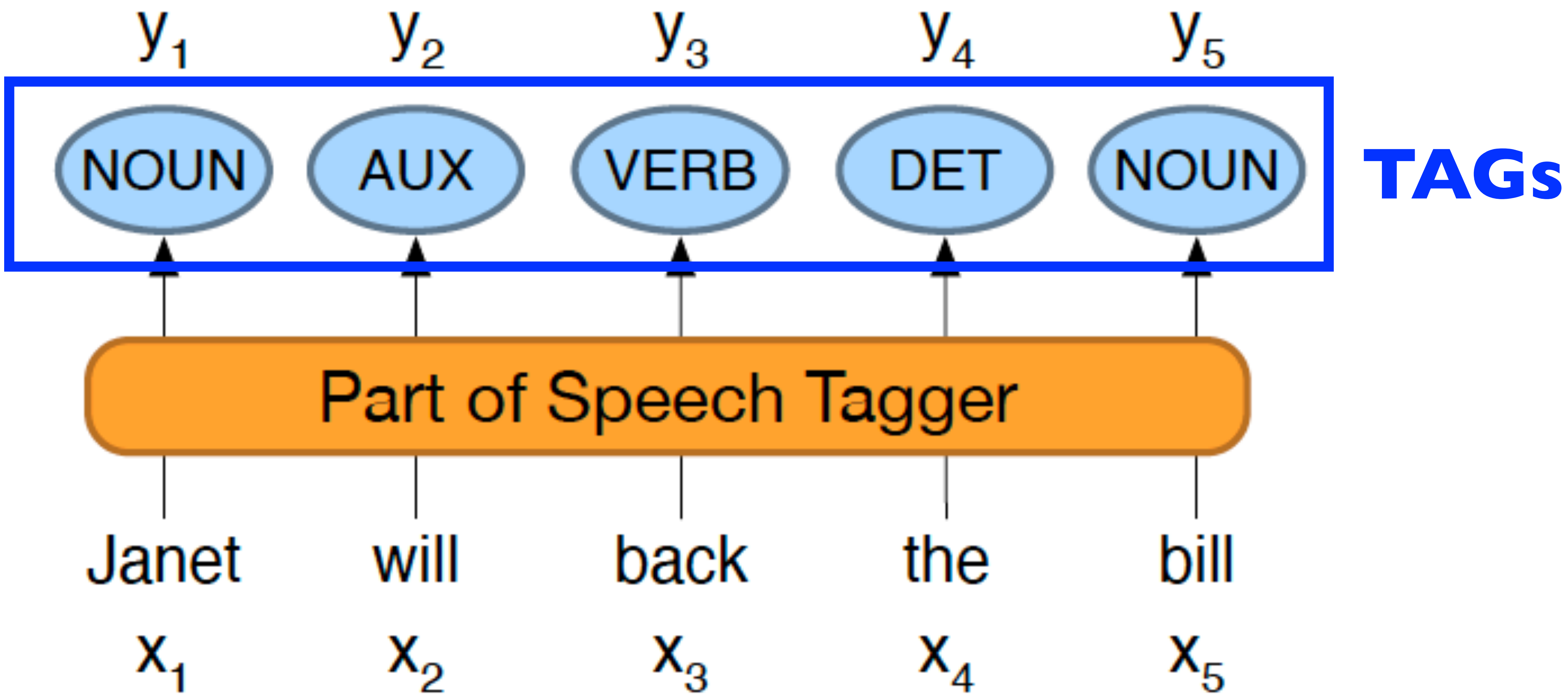
PoS tags in NLP

The PoS tagging task consists in mapping the input words of a sentence in a set of tags selected in a given tag set



PoS tags in NLP

The PoS tagging task consists in mapping the input words of a sentence in a set of tags selected in a given tag set



PoS tagging

To determine the syntactic structure of a sentence (and later its semantics) each word must be recognised as belonging to a grammatical category.
After the tokeniser, **PoS tagger** applies the first analysis.

The activity of the PoS tagger can be seen as:

- disambiguation: given the set of all the possible tags, for each token the tagger discards all the not correct ones

- classification: given a token, it decides the correct category to be associated with that token

PoS tagging and ambiguity

PoS Tagging is a disambiguation task because tokens can be ambiguous: more than one tag can be correctly associated with a token.

PoS tagger has to find the correct tag for the given context.

The problem is that the tagger works each time on a single token, while its association with a specific PoS tag can be decided only observing the tokens before and after it.

In some cases, the ultimate disambiguation can be done only at syntactic level.

PoS tagging and ambiguity

The problem of morphological ambiguity is exacerbated by the fact that **the most frequently used words are the most ambiguous**.

Example: in English the most ambiguous words are *that*, *back*, *down*, *put* and *set*. See how back can be associated with 5 PoS categories:

- earnings growth took a ***back*** seat (Adjective)
- a small building in the ***back*** (Noun)
- a clear majority of senators ***back*** the bill Dave began to ***back*** toward the door enable the country to buy ***back*** about debt (Verb - Verb - Particle)
- I was twenty-one ***back*** then (Adverb)

PoS tagging and ambiguity

How frequent is morphological ambiguity?

Example: the Brown corpus is composed of 1,000,000 tokens corresponding to 39,440 different lemmas.

- 35,340 lemmas can be associated with one single PoS tag anywhere in corpus (89.6 %)
- 4.100(10.4%) can be associated with 2 to 7 different PoS tags.

These 10.4% tokens lead to about 50% of the ambiguity because they are the most frequent tokens in the that corpus.

PoS tagsets in NLP

Within NLP several different PoS tagged corpora and tagsets are available for several different languages.

The inventories of tags meaningfully vary in different projects based on some consideration about:

- the **language** of the corpus
- the **domain** of the corpus
- the presumed **exploitation** of the corpus

PoS tagsets in NLP

Among the most commonly used tagsets are those in:

- the **Penn Treebank** project
- the **Universal Dependencies** project.

In both cases the tagset was developed along with an annotation scheme for the syntax that is applied in the analysis step following PoS tagging.

PoS tags in Penn

The **Penn Treebank** is a private project of the University of Pennsylvania that began in the 1980s. The corpus was first released in 1992.
This large annotated corpus consists of over 4.5 million words of American English.
During the first three-year phase of the project (1989-1992), the entire corpus was annotated for part-of-speech, while more than half of it has been annotated with a syntactic skeleton structure.
The corpus is hosted and marketed by the Linguistic Data Consortium (LDC).

PoS tags in Penn

The rationale behind the **earlier** development of **tagsets** for large corpora (e.g. the Brown Corpus) was to introduce richly structured label sets (with 87 to 190 tags) to provide unique encodings for all word classes of words with unique grammatical behaviour.

The **Penn Treebank tagset** is based on that of the Brown Corpus, which consists of 87 tags.
However, in order to limit sparsity of categories (i.e. categories that occur very infrequently also in a very large corpus) and redundancy, the Brown Corpus tagset was reduced to only 36 tags for words and symbols and 12 tags for punctuation.

PoS tags in Penn

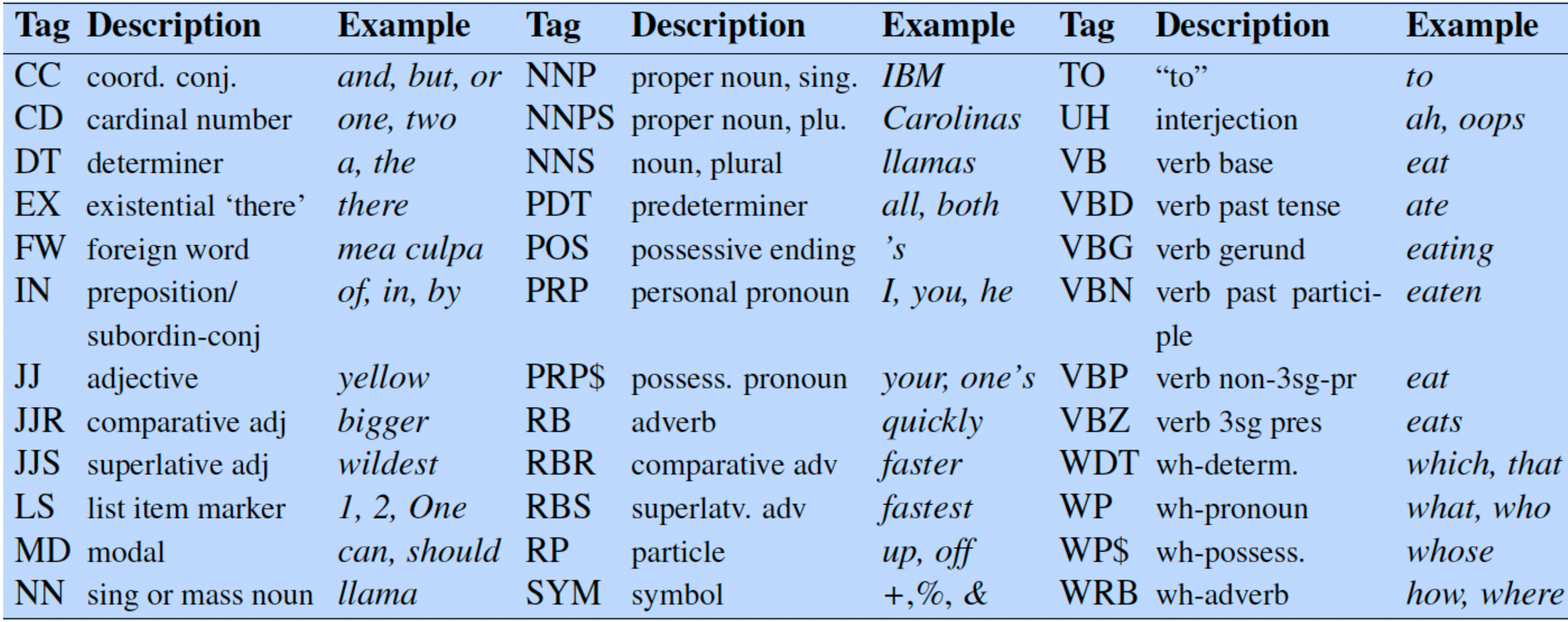
The rationale behind the **earlier** development of **tagsets** for large corpora (e.g. the Brown Corpus) was to introduce richly structured label sets (with 87 to 190 tags) to provide unique encodings for all word classes of words with unique grammati behaviour.

The **Pe** f the Brown Corpus. However, in order to limi gories that occur very infrequently also in a very larg nd redundancy, the Brown Corpus tagset was reduced to only 36 tags for words and symbols and 12 tags for punctuation.

a set of labels is the result of complex theoretical and applied work

PoS tags in Penn

The Penn Treebank labels for words and symbols:



Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	“to”	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential ‘there’	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*’s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one’s*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+,%, &	WRB	wh-adverb	*how, where*

PoS tags in Penn

The Penn Treebank labels for punctuation marks:

#	Pound sign
$	Dollar sign
.	Sentence-final punctuation
,	Comma
:	Colon, semi-colon
(	Left bracket character
)	Right bracket character
"	Straight double quote
`	Left open single quote
"	Left open double quote
'	Right close single quote
"	Right close double quote

Guidelines for Penn PoS tagging

The development of a tagset (and all other forms of annotation) is usually associated with the writing of detailed guidelines that are published together with the annotated data.

Guidelines are essential to make annotation work reproducible! They carefully describe the meaning of each tag and provide several examples of their use in simpler and harder cases.

Penn Treebank annotation guidelines for PoS tagging are available at:

> http://www.cs.cmu.edu/~roni/11761/2017_fall_assignments/hw7-tagguide.pdf

PoS tags in Penn

The tagset used in the Penn Treebank project:

- Contains only **categories, but no features**; this means that a single level of annotation is allowed, which introduces all accepted information granularity
- **Language**: was designed for English, but later adapted to other languages to allow cross-language comparisons
- **Domain**: was designed to deal with financial texts, since an important section of the corpus is drawn from the *Wall Street Journal*
- **Exploitation**: was created to train NLP tools for morphological and syntactic analysis by machine learning.

Raw text: Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Tagged: Pierre / NNP
Vinken / NNP
, / ,
61 / CD
years / NNS
old / JJ
, / ,
will / MD
join / VB
the / DT
board / NN
as / IN
a / DT
nonexecutive / JJ
director / NN
Nov. / NNP
29 / CD
. / .