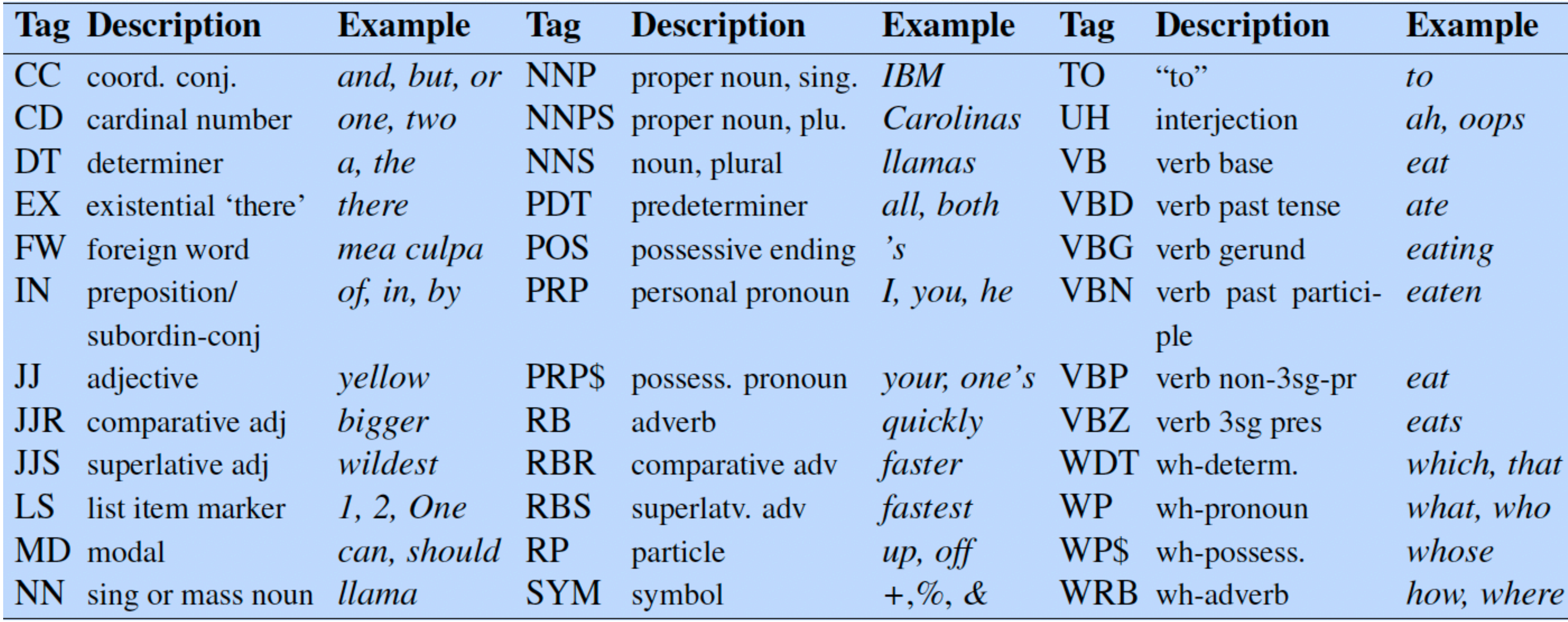
Morphology:
Part of speech tagging
PART 2

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

Cristina Bosco

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

PoS tags in Penn

The tagset used in the Penn Treebank project:

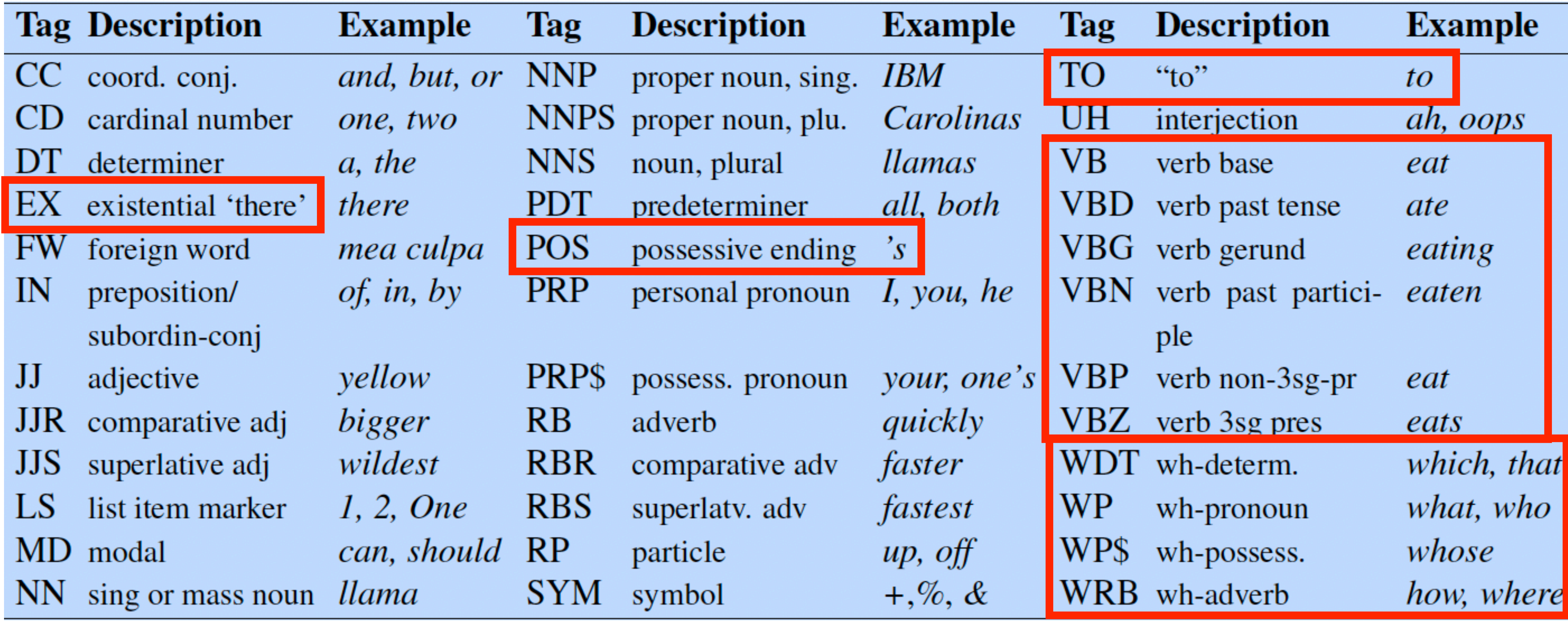
- Contains only **categories, but no features**; this means that a single level of annotation is allowed, which introduces all accepted information granularity
- **Language**: was designed for English, but later adapted to other languages to allow cross-language comparisons
- **Domain**: was designed 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
. / .

PoS tags in Penn

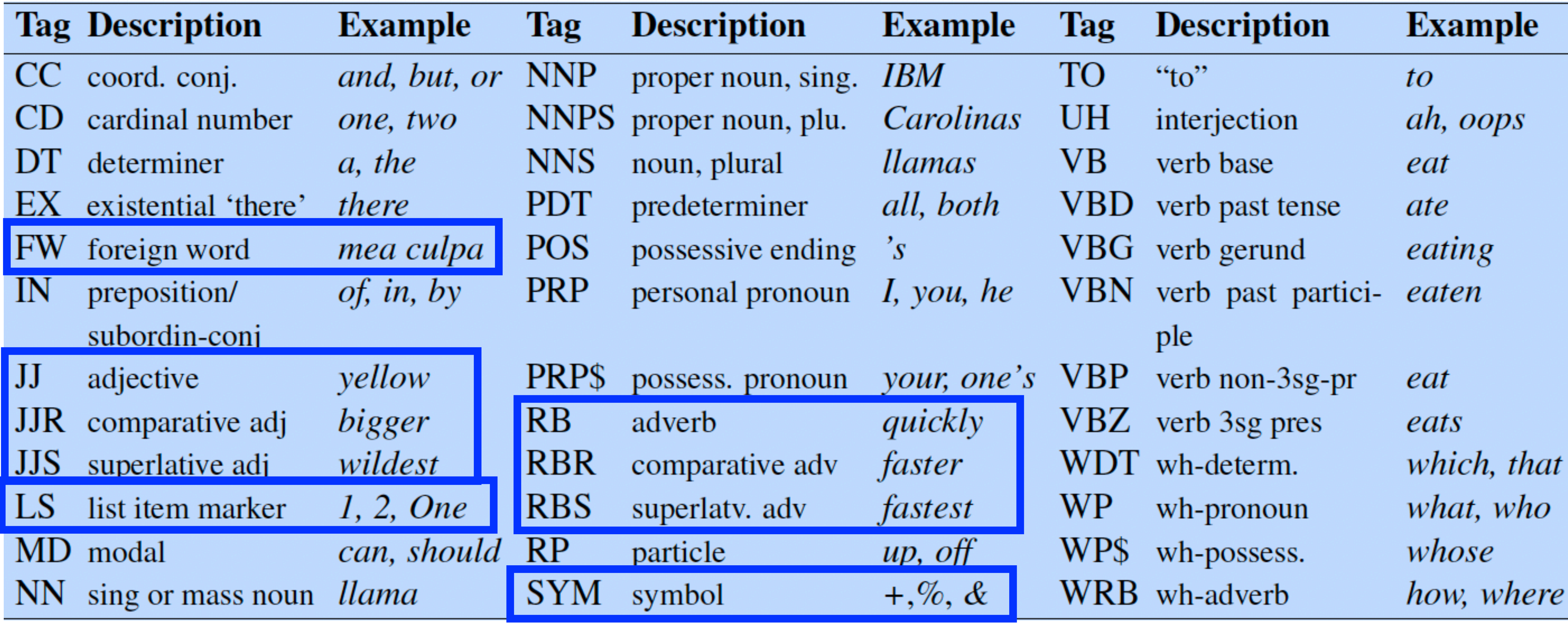
The Penn Treebank labels especially designed for English:



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 especially designed for financial domain:



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 especially designed for financial domain:

#	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

PoS tags in UD

The **Universal Dependencies** is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages.
UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages.

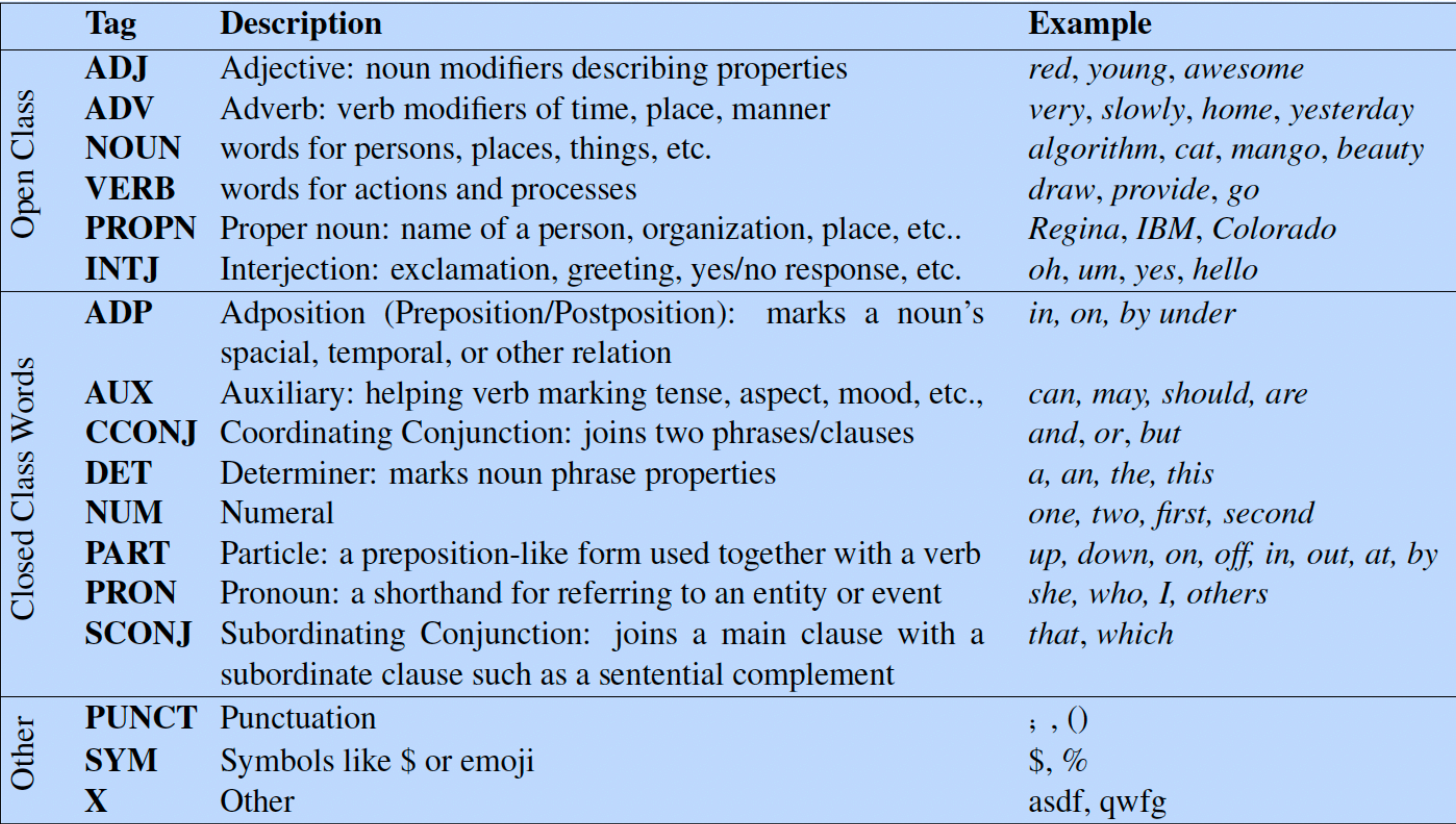
The goal of UD is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.

PoS tags in UD

The **Universal Dependencies** facilitate multilingual parser development, cross-lingual learning, and parsing research.

The annotation scheme is an evolution of Stanford dependencies, Google universal part-of-speech tags, and the Interset interlingua for morphosyntactic tagsets.
The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.

PoS tags in UD



	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red*, *young*, *awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very*, *slowly*, *home*, *yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm*, *cat*, *mango*, *beauty*
	VERB	words for actions and processes	*draw*, *provide*, *go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina*, *IBM*, *Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and*, *or*, *but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that*, *which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

PoS tags in Penn

The guidelines published for tagging tokens according to the UD format are available at this website with a large amount of information and examples:

https://universaldependencies.org/u/pos/index.html

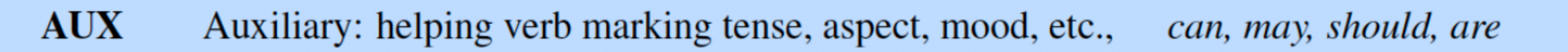
In the sections about different languages and resources, more precise information can be available.

PoS tags in UD

The tagset provided by the Universal Dependencies project includes 17 categories that can be also linked to features.

It introduces the distinction between closed and open classes:

- **Closed** classes have a relatively fixed membership and mostly include function words; also auxiliary and modals are considered as a closed classes



AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*

- **Open** classes can be extended by creating new elements and contain semantically loaded words

PoS tags in UD: CoNLL-U

The **Universal Dependencies** data are released in the 10 column CoNLL-U format and encoded in UTF-8.

The acronym CoNLL stands for **Conference on Computational Natural Language Learning**, the name of the competition (for parsing systems) for which this format was used (for the first time in 2006).

The original format was **CoNLL-X** (in reference to the 10 columns it contains), later it was changed to **CoNLL-U** (in reference to UD).

PoS tags in UD: CoNLL-U

The 10 column CoNLL-U format includes three types of lines:

- **Word** lines containing the annotation of a word/token in 10 fields separated by single tab characters
- **Blank** lines marking sentence boundaries.
- **Comment** lines starting with hash (#)

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

PoS tags in UD

Sentences consist of word lines, which contain the following fields or underscore if not available:

ID: Word index
FORM: Word form or punctuation symbol
LEMMA: Lemma or stem of word form
UPOS: Universal part-of-speech tag
XPOS: Language-specific part-of-speech tag; underscore if not available
FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension

Morphology

HEAD: Head of the current word
DEPREL: Universal dependency relation to the HEAD
DEPS: Enhanced dependency graph
MISC: Any other annotation

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

ID

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

Form

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

```
1   Pierre         Pierre        PROPN    NNP   Number=Sing                   _ _ _
2.  Vinken         Vinken        PROPN    NNP   Number=Sing                   _ _ _
3   ,              ,             PUNCT    ,     _                             _ _ _
4   61             61            NUM      CD    NumType=Card                  _ _ _
5   years          year          NOUN     NNS   Number=Plur                   _ _ _
6   old            old           ADJ      JJ    Degree=Pos                    _ _ _
7   ,              ,             PUNCT    ,     _                             _ _ _
8.  will           will          AUX      MD    VerbForm=Fin                  _ _ _
9   join           join          VERB     VB    VerbForm=Inf                  _ _ _
10 the             the           DET      DT    Definite=Def|PronType=Art     _ _ _
11 board           board.        NOUN     NN    Number=Sing                   _ _ _
12 as              as            ADP      IN    _                             _ _ _
13 a               a             DET.     DT    Definite=Ind|PronType=Art     _ _ _
14 nonexecutive    nonexecutive  ADJ      JJ    Degree=Pos                    _ _ _
15 director        director      NOUN.    NN    Number=Sing                   _ _ _
16 Nov.            November      PROPN    NNP   Abbr=Yes|Number=Sing          _ _ _
17 29              29            NUM      CD    NumType=Card                  _ _ _
18 .               .             PUNCT    .     _                             _ _ _
```

Lemma

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

Upos

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

Xpos

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

1	Pierre	Pierre	PROPN	NNP	Number=Sing	_ _ _
2.	Vinken	Vinken	PROPN	NNP	Number=Sing	_ _ _
3	,	,	PUNCT	,	_	_ _ _
4	61	61	NUM	CD	NumType=Card	_ _ _
5	years	year	NOUN	NNS	Number=Plur	_ _ _
6	old	old	ADJ	JJ	Degree=Pos	_ _ _
7	,	,	PUNCT	,	_	_ _ _
8.	will	will	AUX	MD	VerbForm=Fin	_ _ _
9	join	join	VERB	VB	VerbForm=Inf	_ _ _
10	the	the	DET	DT	Definite=Def|PronType=Art	_ _ _
11	board	board.	NOUN	NN	Number=Sing	_ _ _
12	as	as	ADP	IN	_	_ _ _
13	a	a	DET.	DT	Definite=Ind|PronType=Art	_ _ _
14	nonexecutive	nonexecutive	ADJ	JJ	Degree=Pos	_ _ _
15	director	director	NOUN.	NN	Number=Sing	_ _ _
16	Nov.	November	PROPN	NNP	Abbr=Yes|Number=Sing	_ _ _
17	29	29	NUM	CD	NumType=Card	_ _ _
18	.	.	PUNCT	.	_	_ _ _

Features

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

```
1   Pierre         Pierre         PROPN    NNP   Number=Sing                      _ _ _
2.  Vinken         Vinken         PROPN    NNP   Number=Sing                      _ _ _
3   ,              ,              PUNCT    ,     _                                _ _ _
4   61             61             NUM      CD    NumType=Card                     _ _ _
5   years          year           NOUN     NNS   Number=Plur                      _ _ _
6   old            old            ADJ      JJ    Degree=Pos                       _ _ _
7   ,              ,              PUNCT    ,     _                                _ _ _
8.  will           will           AUX      MD    VerbForm=Fin                     _ _ _
9   join           join           VERB     VB    VerbForm=Inf                     _ _ _
10  the            the            DET      DT    Definite=Def|PronType=Art        _ _ _
11  board          board.         NOUN     NN    Number=Sing                      _ _ _
12  as             as             ADP      IN    _                                _ _ _
13  a              a              DET.     DT    Definite=Ind|PronType=Art        _ _ _
14  nonexecutive   nonexecutive   ADJ      JJ    Degree=Pos                       _ _ _
15  director       director       NOUN.    NN    Number=Sing                      _ _ _
16  Nov.           November       PROPN    NNP   Abbr=Yes|Number=Sing             _ _ _
17  29             29             NUM      CD    NumType=Card                     _ _ _
18  .              .              PUNCT    .     _                                _ _ _
```

Upos

Features

How much UD is universal?

ITALIAN

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PART – PRON – PROPN – PUNCT – SCONJ – SYM –VERB – X

FRENCH

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PRON – PROPN – PUNCT – SCONJ – SYM –VERB – X

ENGLISH

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PART – PRON – PROPN – PUNCT – SCONJ – SYM –VERB – X

SPANISH

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PART – PRON – PROPN – PUNCT – SCONJ – SYM –VERB – X

How much UD is universal?

ITALIAN

Clitic – Definite – Degree – Foreign – Gender – Mood – Number – NumType – Person – Polarity – Poss – PronType – Tense – VerbForm

FRENCH

Definite – Foreign – Gender – Mood – Number – Number[psor] – NumType – Person – Person[psor] – Polarity – Poss – PronType – Reflex – Tense – Typo – VerbForm

ENGLISH

Abbr – Case – Definite – Degree – Gender – Mood – Number – NumForm – NumType – Person – Polarity – Poss – PronType – Reflex – Tense – Typo – VerbForm – Voice

SPANISH

AdvType – Case – Definite – Degree – Foreign – Gender – Mood – Number – Number[psor] – NumForm – NumType – Person – Polarity – Polite – Poss – PrepCase – PronType – PunctSide – PunctType – Reflex – Tense – Typo – VerbForm

How much UD is universal?

ITALIAN

Clitic – [✓ Definite] – Degree – Foreign – [✓ Gender] – [✓ Mood] – [✓ Number] – [✓ NumType] – [✓ Person] – [✓ Polarity] – [✓ Poss] – [✓ PronType] – [✓ Tense] – [✓ VerbForm]

FRENCH

[✓ Definite] – Foreign – [✓ Gender] – [✓ Mood] – [✓ Number] – Number[psor] – [✓ NumType] – [✓ Person] – Person[psor] – [✓ Polarity] – [✓ Poss] – [✓ PronType] – Reflex – [✓ Tense] – Typo – [✓ VerbForm]

ENGLISH

Abbr – Case – [✓ Definite] – Degree – [✓ Gender] – [✓ Mood] – [✓ Number] – NumForm – [✓ NumType] – [✓ Person] – [✓ Polarity] – [✓ Poss] – [✓ PronType] – Reflex – [✓ Tense] – Typo – [✓ VerbForm] – Voice

SPANISH

AdvType – Case – [✓ Definite] – Degree – Foreign – [✓ Gender] – [✓ Mood] – [✓ Number] – Number[psor] – NumForm – [✓ NumType] – [✓ Person] – [✓ Polarity] – Polite – [✓ Poss] – PrepCase – [✓ PronType] – PunctSide – PunctType – Reflex – [✓ Tense] – Typo – [✓ VerbForm]

PoS tags

Finer classifications not considered in the tasgsets of the Penn Treebank or in the categories of UD include for example:

- The distinction between **count nouns**, which can occur in the singular and plural (*goat/goats, relationship/relationships*) and can be counted (*one goat, two goats*) / and **mass nouns**, which are used when something is conceptualized as a homogeneous group (*snow, salt, communism, water*)

PoS tags

- **Directional or locative adverbs** (*home, here, downhill*), which specify the direction or location of some action / **degree adverbs** (*extremely, very, somewhat*), which specify the extent of some action, process, or property / **manner adverbs** (*slowly, slinkily, delicately*), which specify the manner of some action or process / **temporal adverbs** (*yesterday, Monday*), which specify the time that some action or event took place. i

PoS tagging RECAP

- Pos-tagging consists of **assigning a tag** (a grammatical category) to a token
- It is **applied to individual tokens** (the context consisting of the surrounding tokens is not taken into account)
- **Different tag sets** can be used (e.g. Penn Treebank and UD)
- It is applied **after tokenization**: Tokenization is necessary to isolate the elements to be tagged
- It is applied **before parsing**: PoS tagging provides information about morphological categories whose behaviour is necessary to know to identify the higher-level syntactic structure

POS TAGGING

It makes explicit a lot of information about each **single word**:

- The tagger observes indeed each word **out of the context** where it occurs, not considering the syntactic links that it can have with some other word of the sentence

il	**ARTICLE**
cane	
dormiva	
nel	
giardino	

POS TAGGING

It makes explicit a lot of information about each **single word**:

- The tagger observes indeed each word out of the context where it occurs, not considering the syntactic links that it can have with some other word of the sentence

il	**ARTICLE**
cane	**NOUN**
dormiva	
nel	
giardino	

POS TAGGING

It makes explicit a lot of information about each **single word**:

- The tagger observes indeed each word out of the context where it occurs, not considering the syntactic links that it can have with some other word of the sentence

il	**ARTICLE**
cane	**NOUN**
dormiva	**VERB**
nel	
giardino	

POS TAGGING

It makes explicit a lot of information about each **single word**:

- The tagger observes indeed each word out of the context where it occurs, not considering the syntactic links that it can have with some other word of the sentence

il	**ARTICLE**
cane	**NOUN**
dormiva	**VERB**
nel	**PREPOSITION + ARTICLE**
giardino	

POS TAGGING

It makes explicit a lot of information about each **single word**:

- The tagger observes indeed each word out of the context where it occurs, not considering the syntactic links that it can have with some other word of the sentence

il	**ARTICLE**
cane	**NOUN**
dormiva	**VERB**
nel	**PREPOSITION + ARTICLE**
giardino	**NOUN**

POS TAGGING

The morphological information is crucial for the following syntactic analysis:

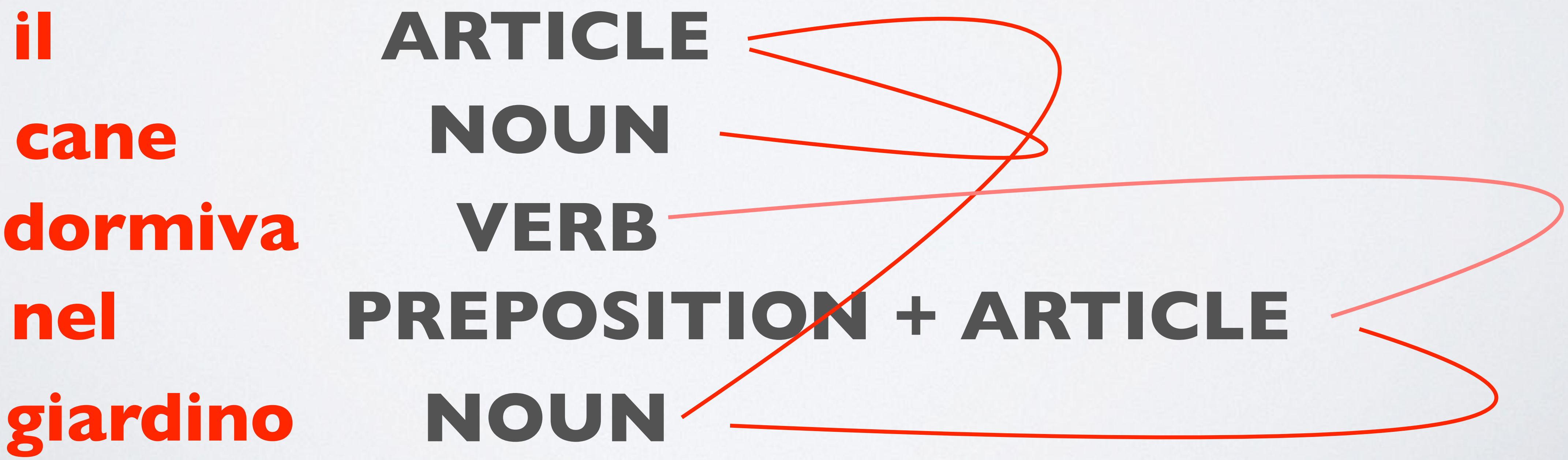
- To recognise the grammatical category of a word W means to know with which other categories of words W can be syntactically related whitin the sentence, and also the categories of words that cannot be related with it

il	**ARTICLE**
cane	**NOUN**
dormiva	**VERB**
nel	**PREPOSITION + ARTICLE**
giardino	**NOUN**

POS TAGGING

The morphological information is crucial for the following syntactic analysis:

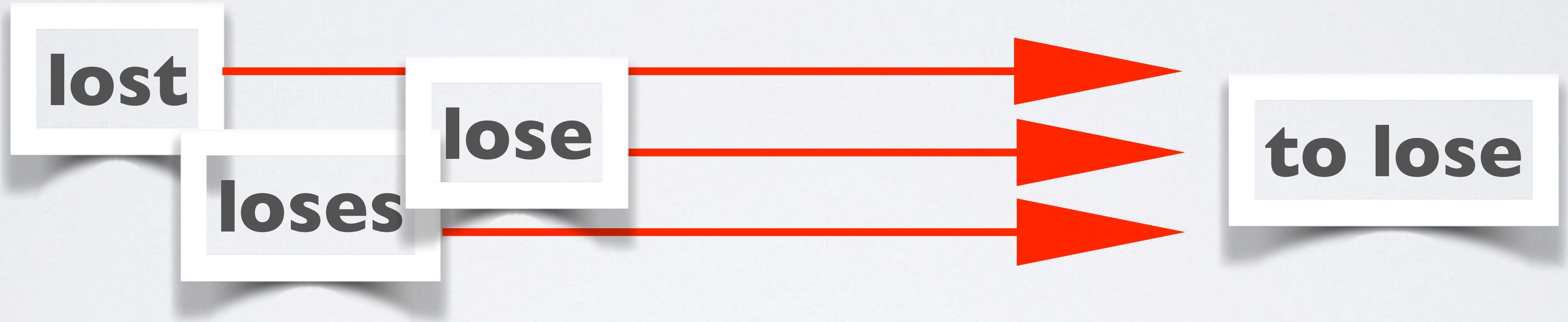
- To recognise the grammatical category of a word W means to know with which other categories of words W can be syntactically related whitin the sentence, and also the categories of words that cannot be related with it



il	ARTICLE
cane	NOUN
dormiva	VERB
nel	PREPOSITION + ARTICLE
giardino	NOUN

POS TAGGING - LEMMATIZATION AND STEMMING

- Lemmatization consists in sorting words by grouping inflected or variant forms of the same word



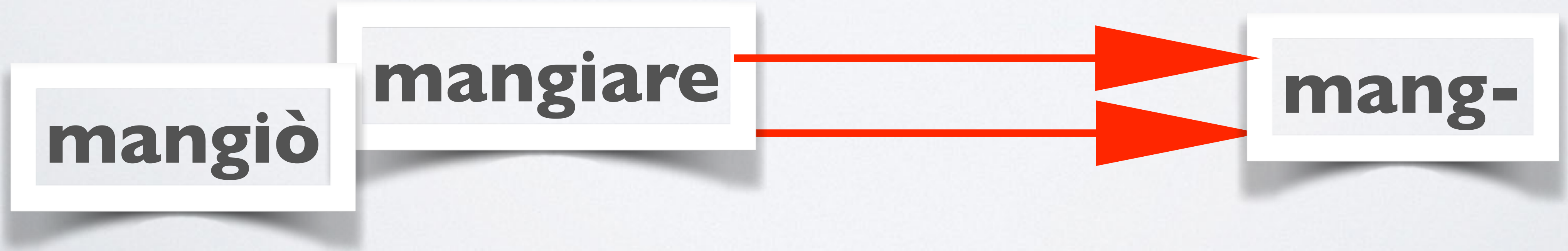
lost

lose

loses

to lose

- Stemming is reducing inflected (or sometimes derived) words to their word stem, base or root form



mangiò

mangiare

mang-

Stemming and lemmatisation

Stemming more specifically refers to a process of **truncating the ends of words**, their inflective part.

For weakly inflected languages this process produces a correct result in most cases and often includes the removal of derivational affixes.
For highly inflected languages and languages with irregular base forms (such as Italian), stemming cannot produce correct forms and lemmatisation is usually applied.

Stemming and lemmatisation

Example:

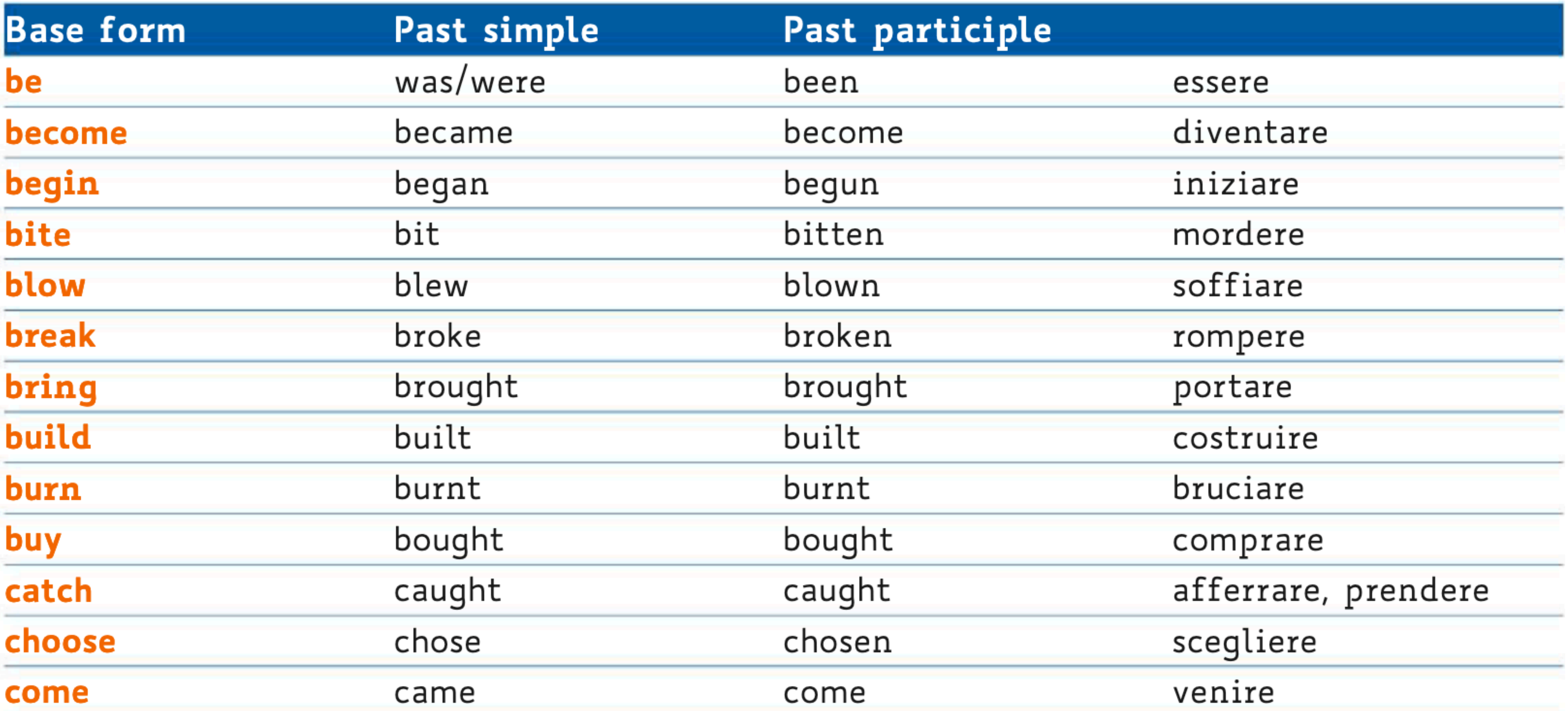
Italian irregular verb for which more than one stem is used

andare (to go):

io andai (I went) same stem of the lemma

io vado (I go) different stem of the lemma

English



Base form	Past simple	Past participle	
be	was/were	been	essere
become	became	become	diventare
begin	began	begun	iniziare
bite	bit	bitten	mordere
blow	blew	blown	soffiare
break	broke	broken	rompere
bring	brought	brought	portare
build	built	built	costruire
burn	burnt	burnt	bruciare
buy	bought	bought	comprare
catch	caught	caught	afferrare, prendere
choose	chose	chosen	scegliere
come	came	come	venire

Stemming and lemmatisation

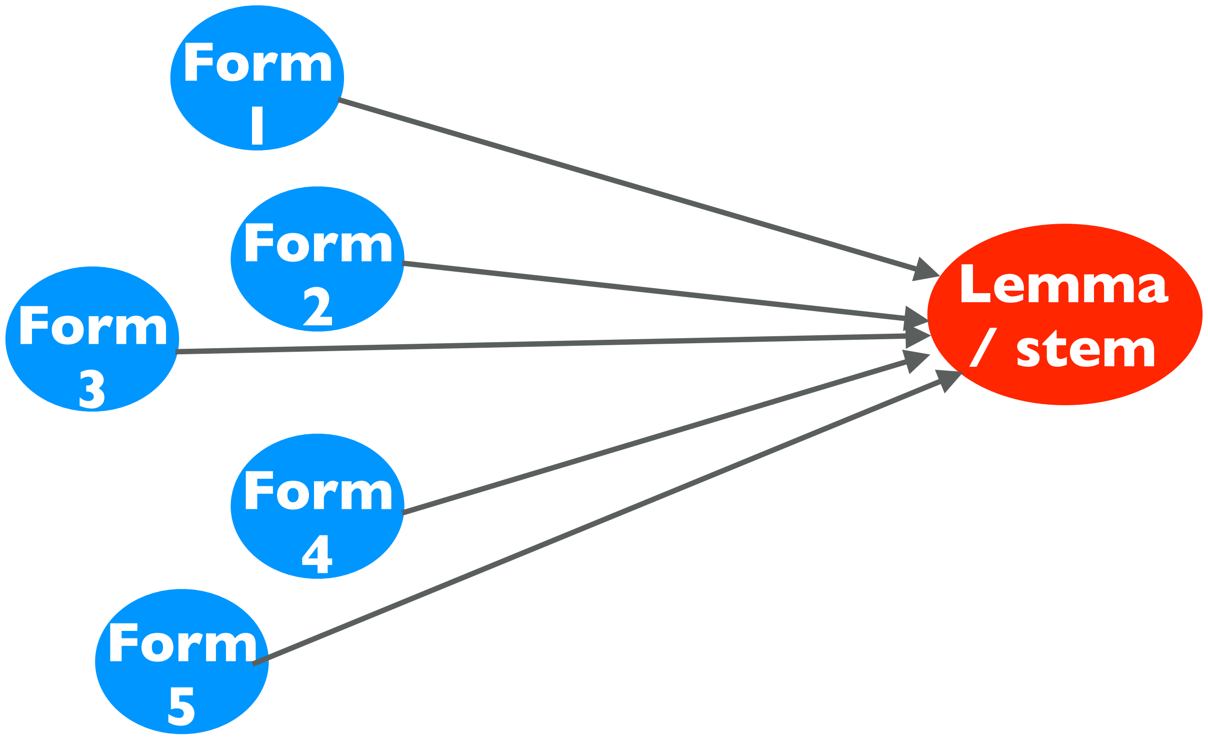
Lemmatization usually refers to return the base or dictionary form of a word, which is known as the *lemma*.
It requires more complex analysis algorithms and linguistic resources with respect to stemming. For example, for Italian the most commonly used resource is Morph-it (by Marco Baroni: https://github.com/giodegas/morphit-lemmatizer/blob/master/master/morph-it_048.txt)

Stemming also most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

Stemming and lemmatisation

The goal of both stemming and lemmatization is to reduce the variety of inflectional forms of a word to a common **base form.**

In a computational perspective, this means to reduce the sparseness of the linguistic data.



POS TAGGING - LEMMATIZATION AND STEMMING

- The main effects of lemmatization and stemming are:

- to reduce lexicon sparseness and vocabulary size

- to make words ready to be searched in (semantic) lexica, dictionaries or other similar resources

POS TAGGING - LEMMATIZATION AND STEMMING

- Lemmatization and stemming can be especially difficult when words are newly created or modified by users like in **social media** texts

stracucciolino

(meaning is around *small and very tender puppy*)

beeeeeeeeeeeello

(*beautifuuuuuuuul*)

manif

manifestation (French)
(*manifestation*)

tvb

ti voglio bene
(*I love you*)

spt

soprattutto
(*in particular*)