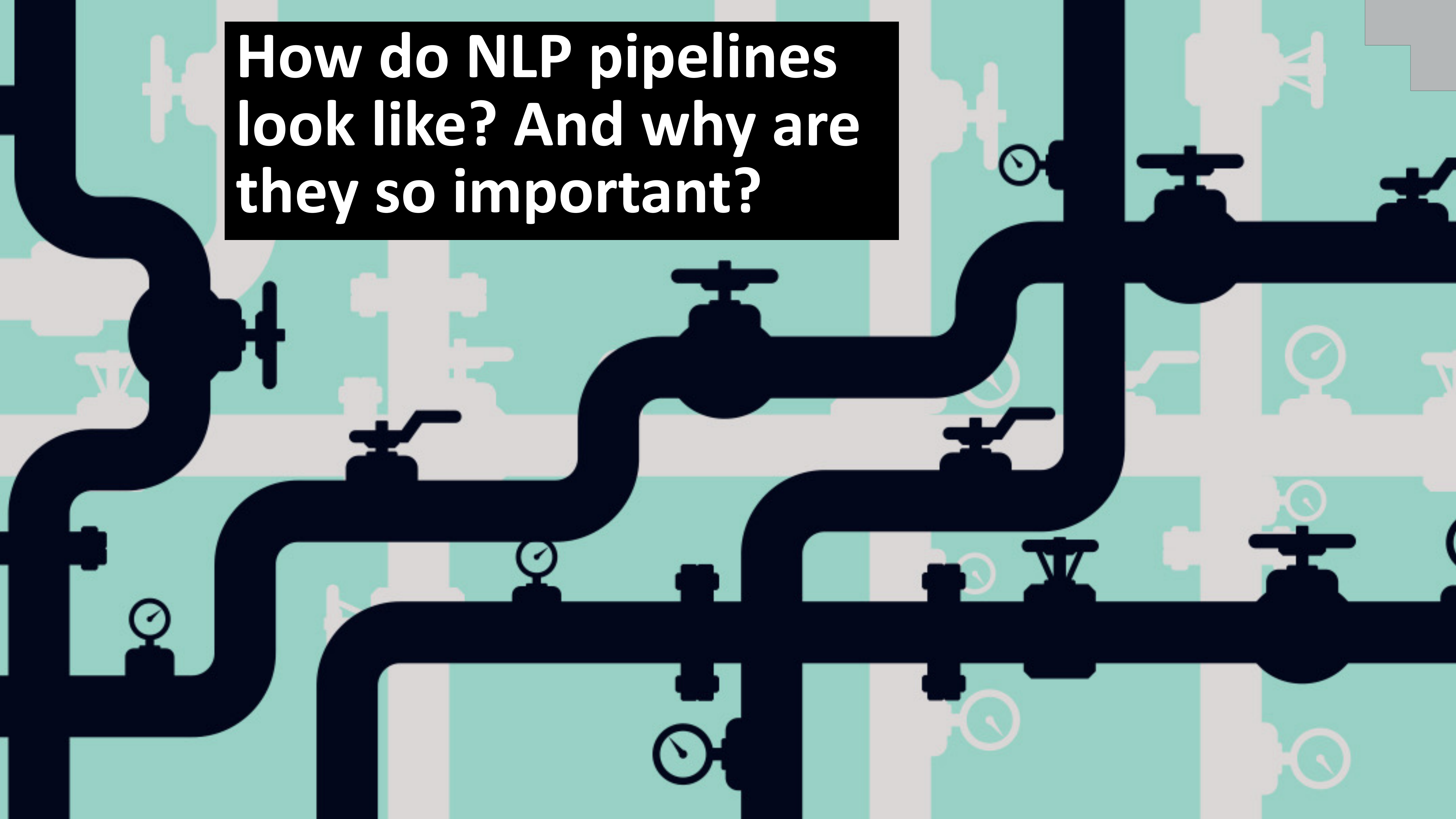
NLP Pipelines

Goals

1. Appreciating the boundaries of a typical NLP pipeline
2. Familiarising with an NLP pipeline’s individual components



How do NLP pipelines look like? And why are they so important?

Why do we use NLP pipelines? (1/2)

Typically, NLP applications do not use a text corpus in its ‘raw’ format.

Instead, they use a transformed text corpus achieved by processing the ‘raw’ text corpus through an NLP pipeline.

Using a transformed text corpus presents substantial advantages; for example:

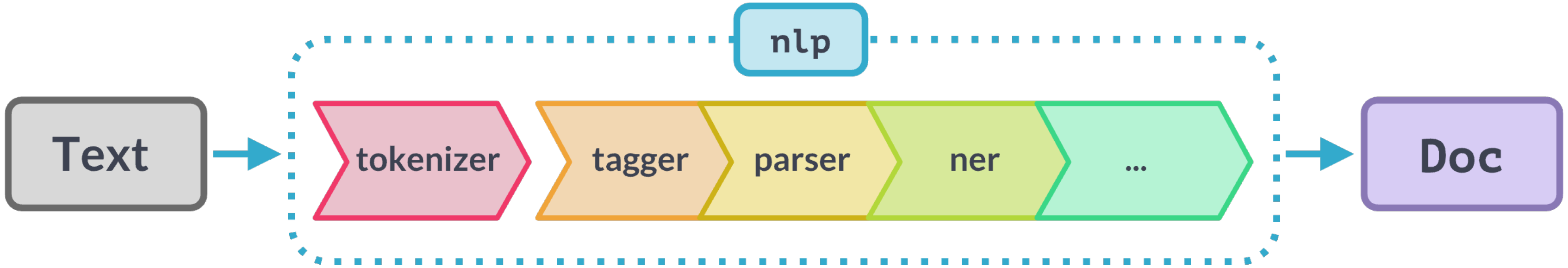
- **Standardisation benefits**: words are tagged with various meta-data (e.g., we know when the term ‘apple’ refers to the company and when it relates to a/the fruit). Hence, we can rely on meta-data to filter in only the terms that are important for the analysis (e.g., words associated with products such as ‘iOS’).
- **Reduced data dimensionalit**y: words can be replaced with lemmas (‘had’ → ‘have’) and stop-words removed (e.g., ‘and’).

Why do we use NLP pipelines? (2/2)

- **Manageable input**: ‘long’ documents can be broken into more manageable segments, such as paragraphs, sentences, or tokens.
- **Cleaner inputs**: stop-words, which bring limited information, can be easily removed from the dataset.

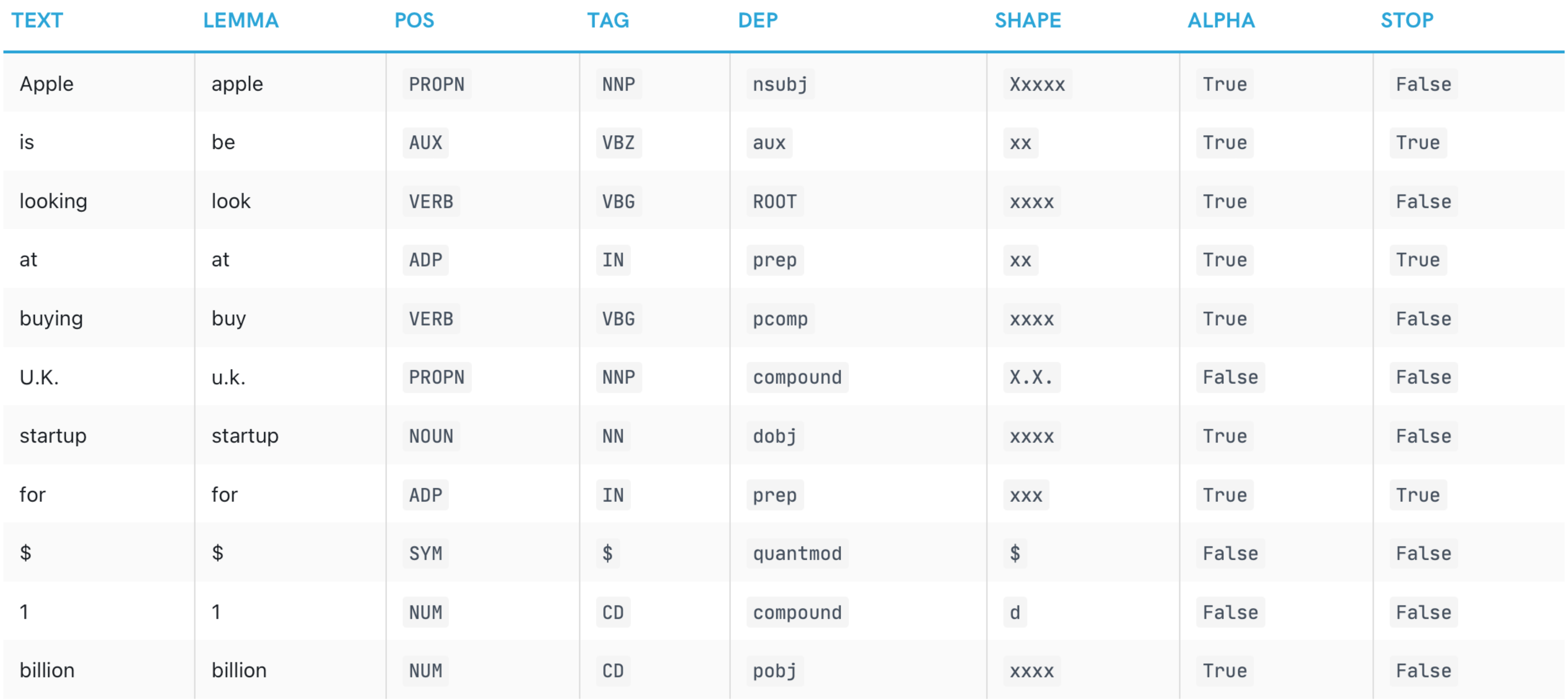
How do NLP pipelines look like?

A typical NLP pipeline.



Source: spaCy documentation.

An NLP pipeline’s outcome.



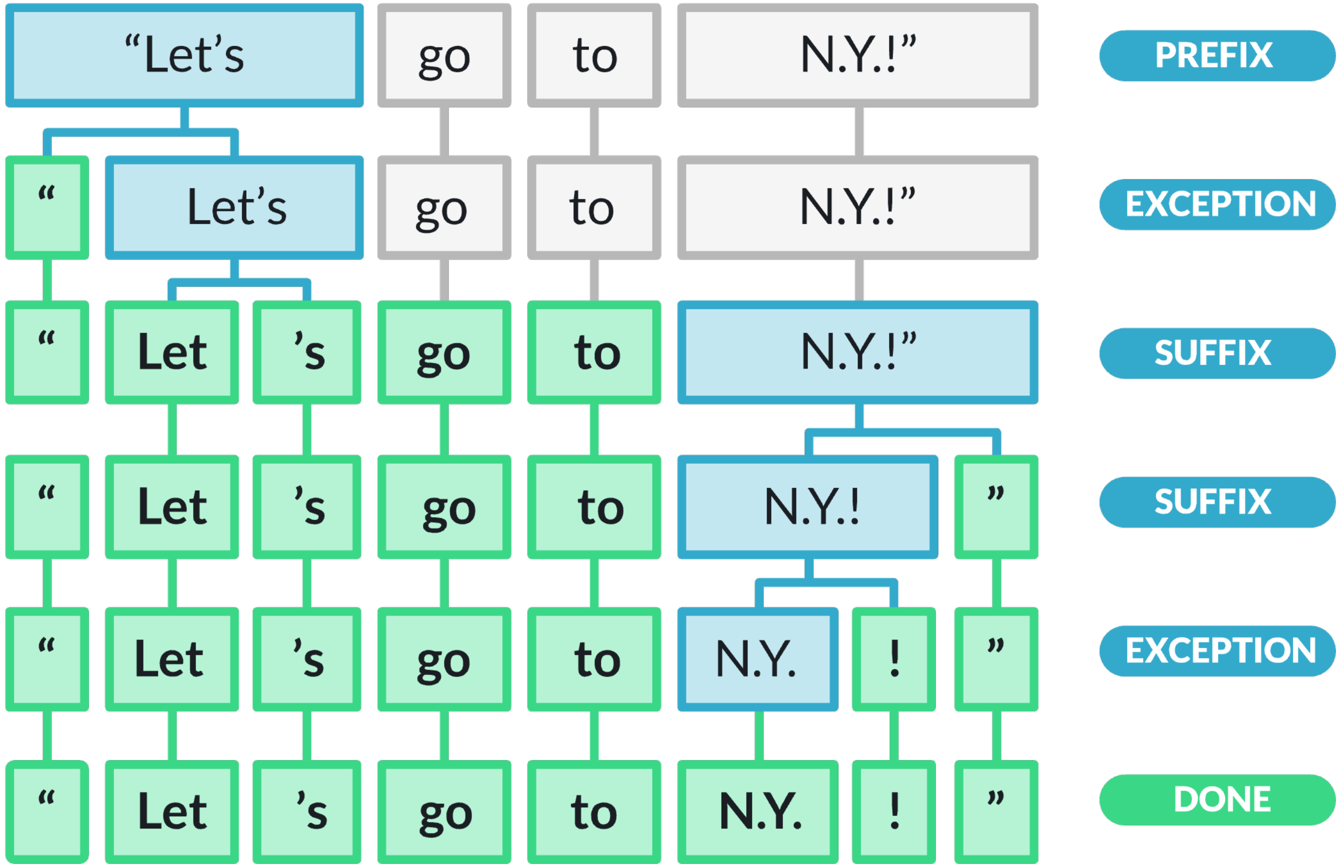
TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

Source: spaCy documentation.

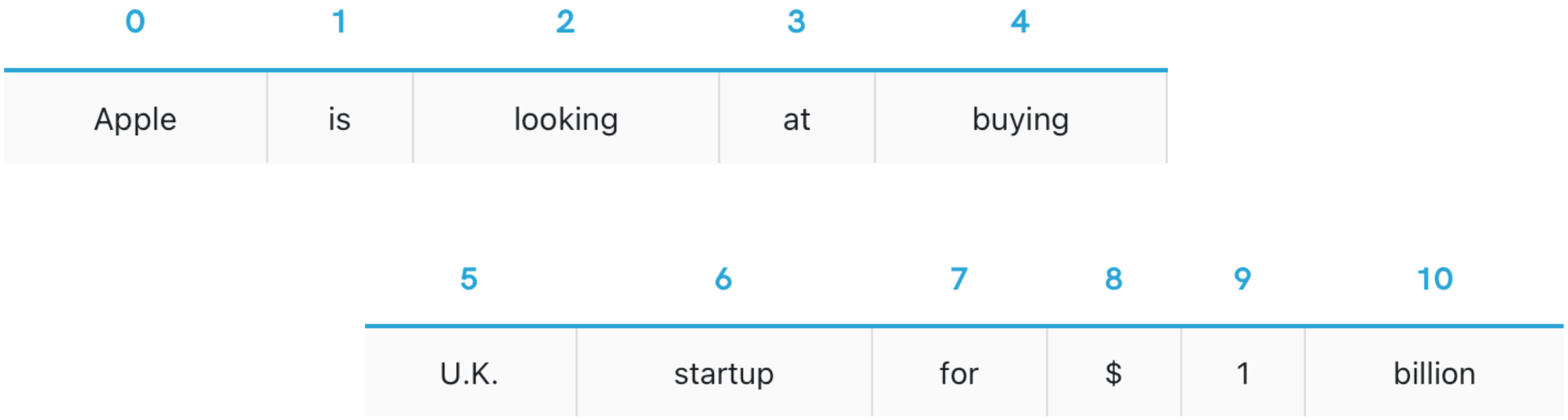
Tokenisation

In a nutshell

A tokeniser splits a text into meaningful segments, called *tokens*.



Source: spaCy documentation.



0	1	2	3	4
Apple	is	looking	at	buying

5	6	7	8	9	10
U.K.	startup	for	$	1	billion

Source: spaCy documentation.

Tagging

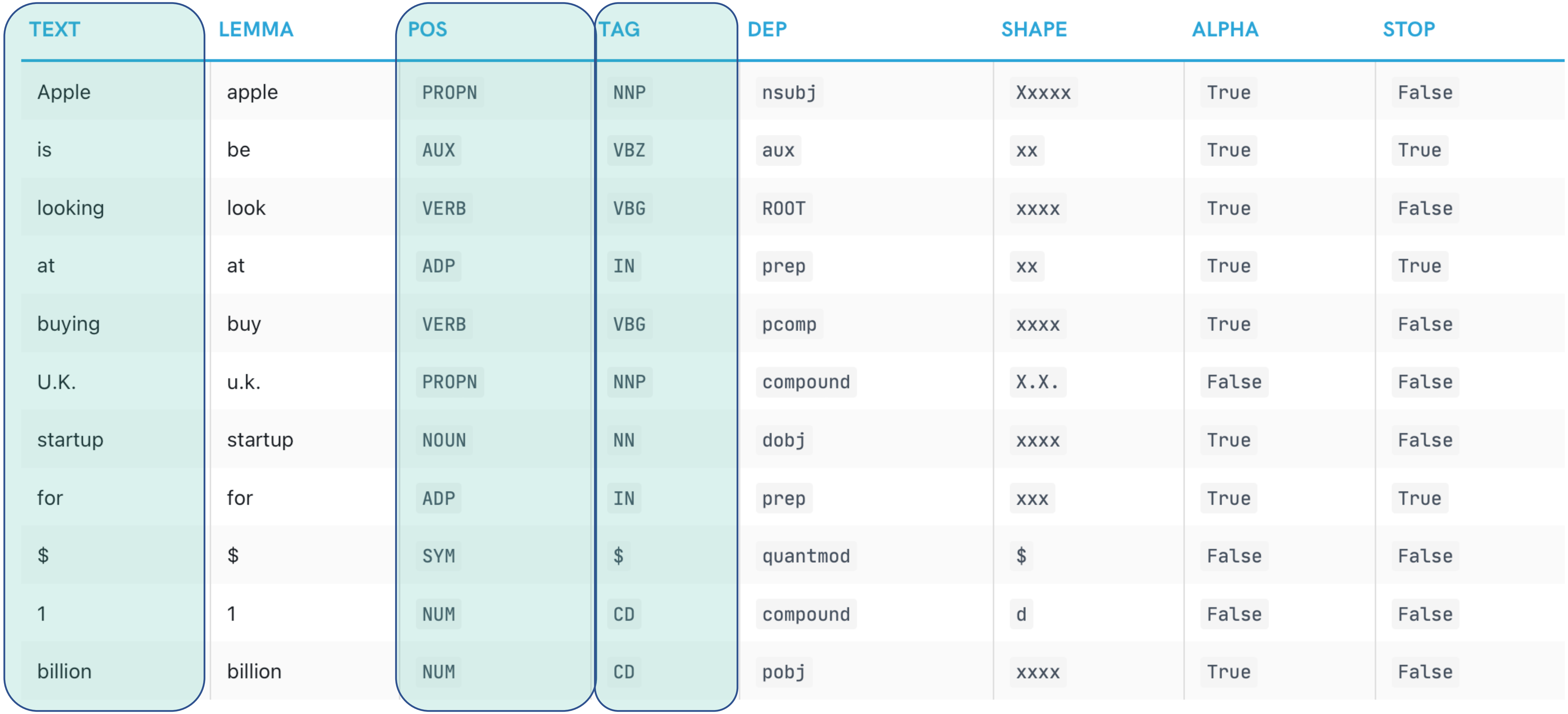
In a nutshell

Once a corpus of text has been tokenised, statistical language models kick in to provide ‘part-of-speech’ tags.

As a result, tags are associated with the individual words.

For example, ‘Apple’ (the company) will be tagged as a ‘proper name’, while ‘apple’ (the fruit) will be classified as a ‘noun’.

Typically, NLP libraries use Universal POS tags.



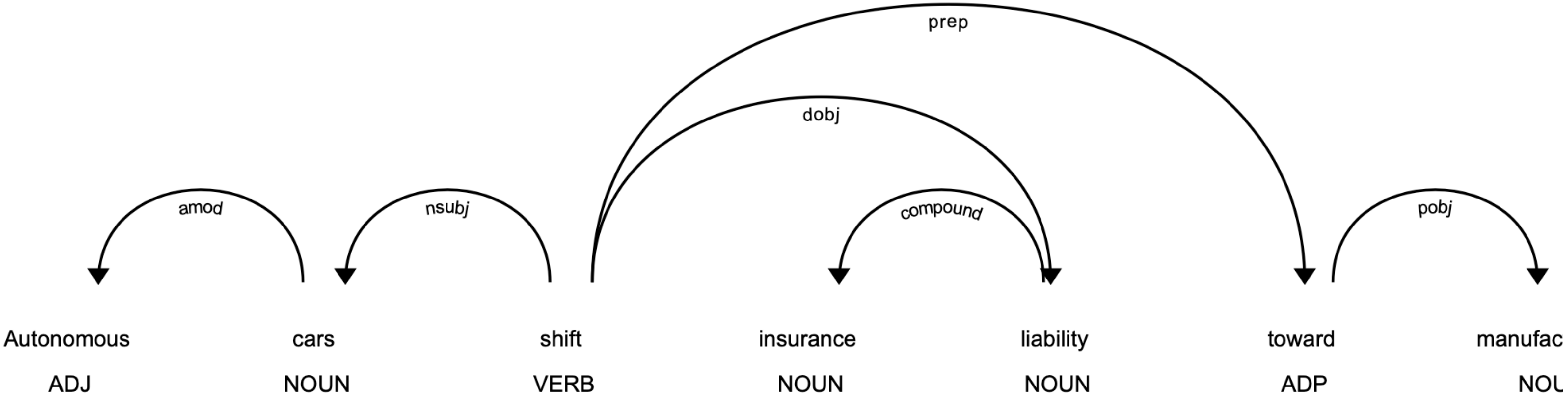
TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

Source: spaCy documentation.

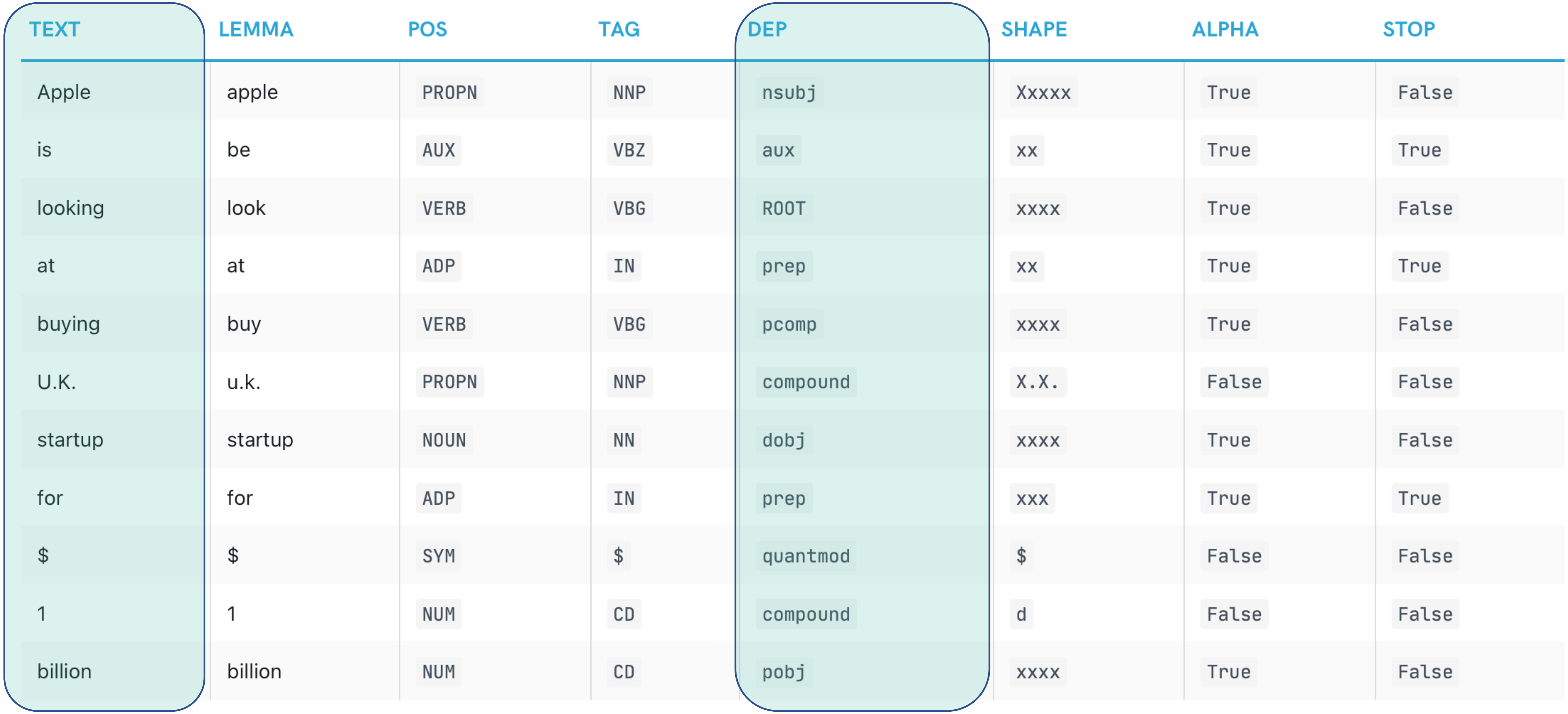
Parsing

In a nutshell

A dependency parser annotates tokens with syntactic dependency labels.



Source: spaCy documentation.



TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

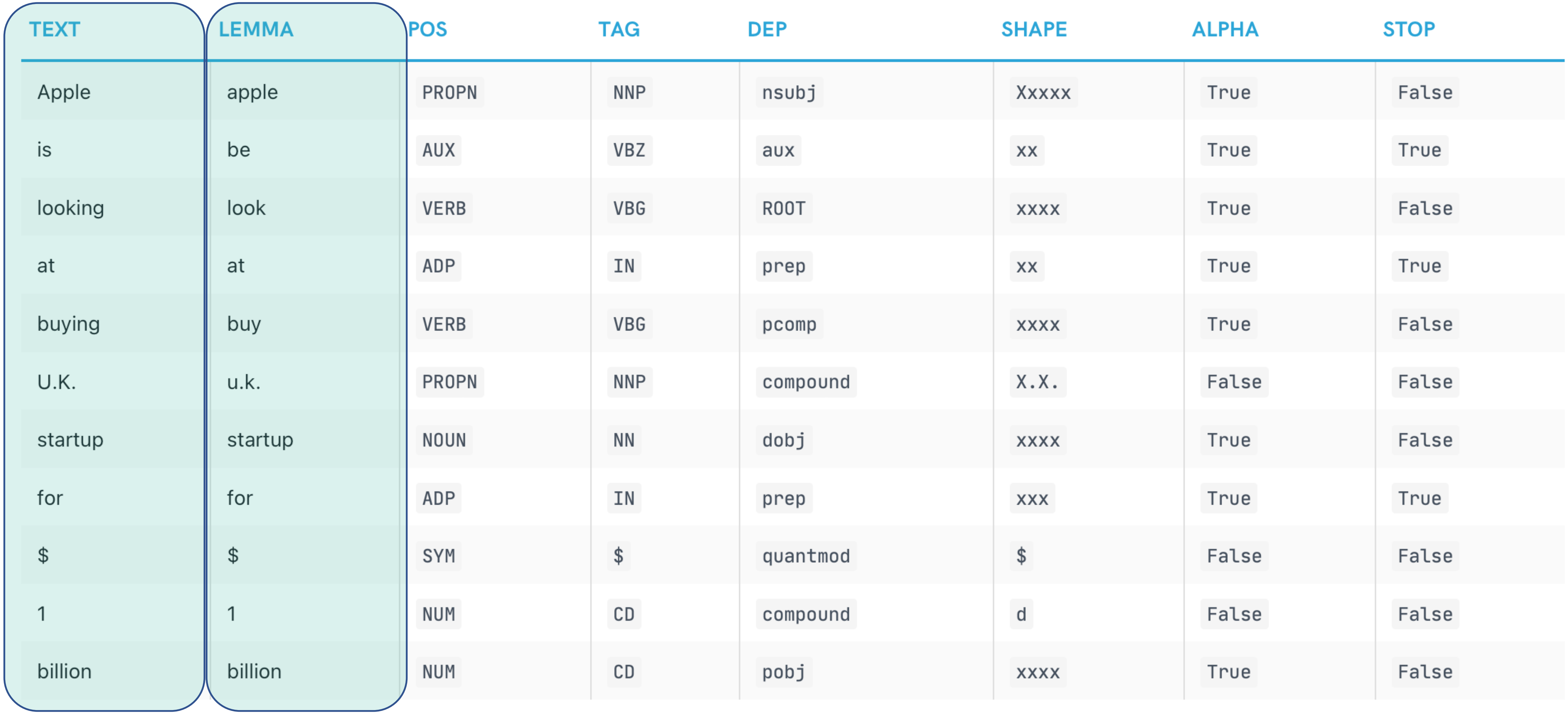
Source: spaCy documentation.

Lemmatization

In a nutshell

A lemmatiser groups the inflected forms of a words around the base form.

For example, the token ‘walking’ is replace with the base form ‘walk’.



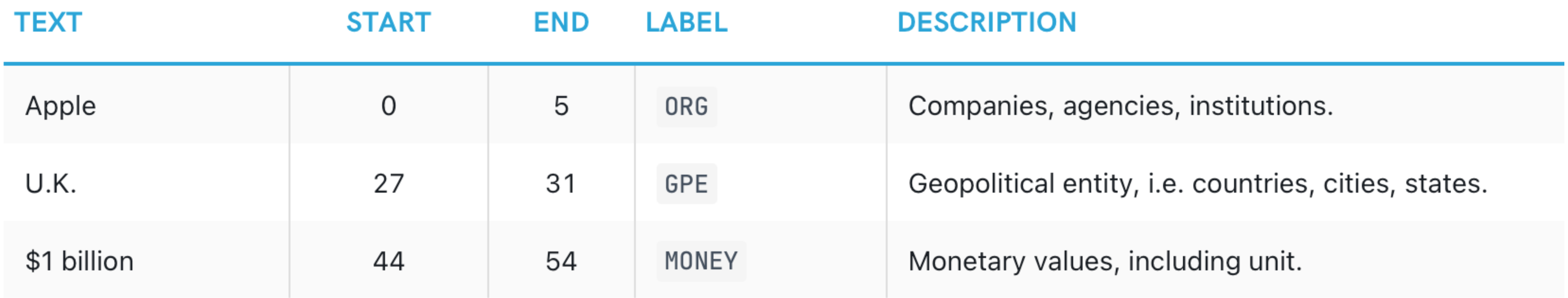
TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

Source: spaCy documentation.

Named-Entity-Recognition

In a nutshell

An entity recognizer associates tokens with types of entities, such as ‘person’, ‘product’, ‘firm’, ‘country’.



TEXT	START	END	LABEL	DESCRIPTION
Apple	0	5	`ORG`	Companies, agencies, institutions.
U.K.	27	31	`GPE`	Geopolitical entity, i.e. countries, cities, states.
$1 billion	44	54	`MONEY`	Monetary values, including unit.

Source: spaCy documentation.

Wrap-up

Main points

- NLP pipelines help analysts improve the quality of their inputs for topic modelling, sentiment analysis, etc.
- A typical NLP pipeline comprises tokenisation, lemmatisation, tagging, and NER capabilities.