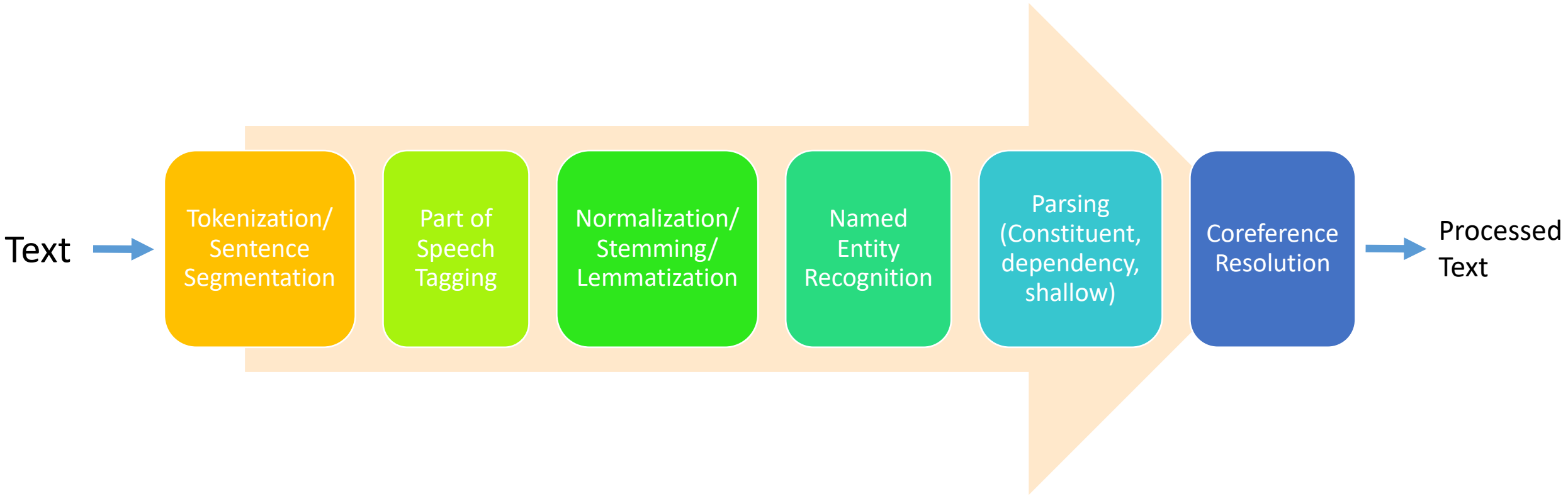
Lecture 1: Text Mining I

Jiaheng Xie
Department of Management Information Systems
Eller College of Management
University of Arizona

Learning Objectives

- A run-through example
- Understand tokenization, part-of-speech tagging, normalization/stemming/lemmatization, named entity recognition, parsing, and coreference resolution

Typical Text Processing Pipeline



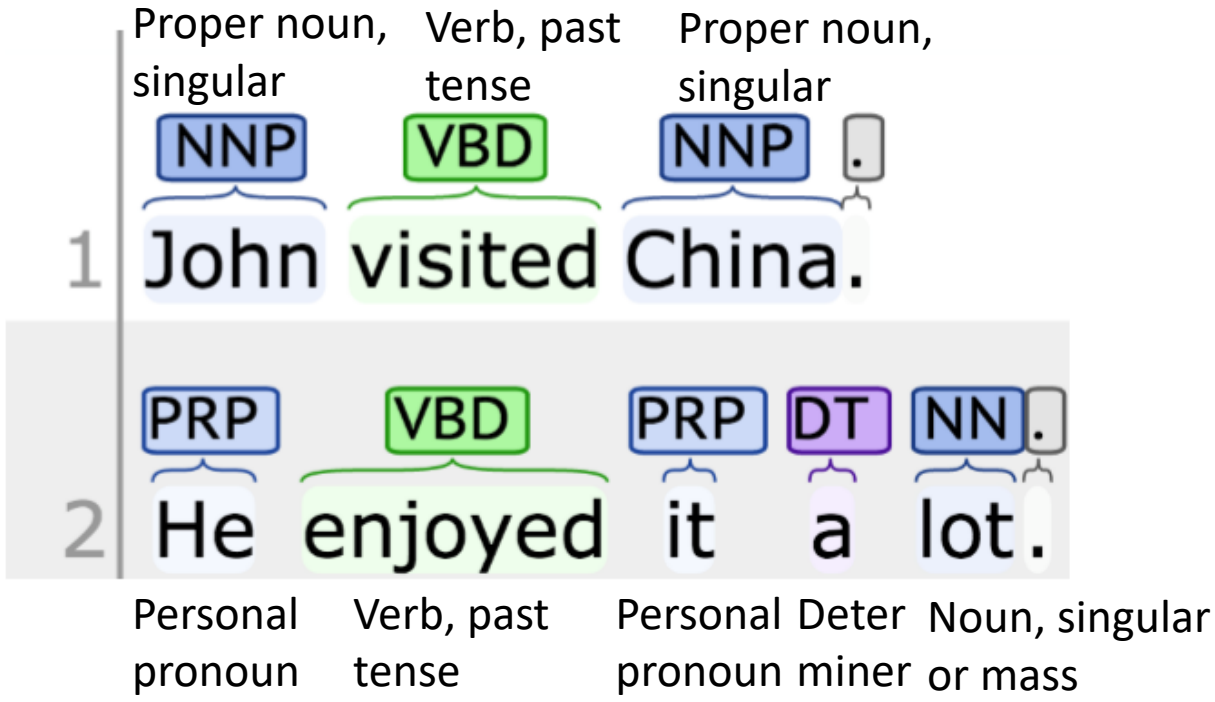
A run-through example

“John visited China. He enjoyed it a lot.”

A run-through example

After tokenization, sentence segmentation, and POS tagging

"John visited China. He enjoyed it a lot."



A run-through example

After lemmatization/ normalization

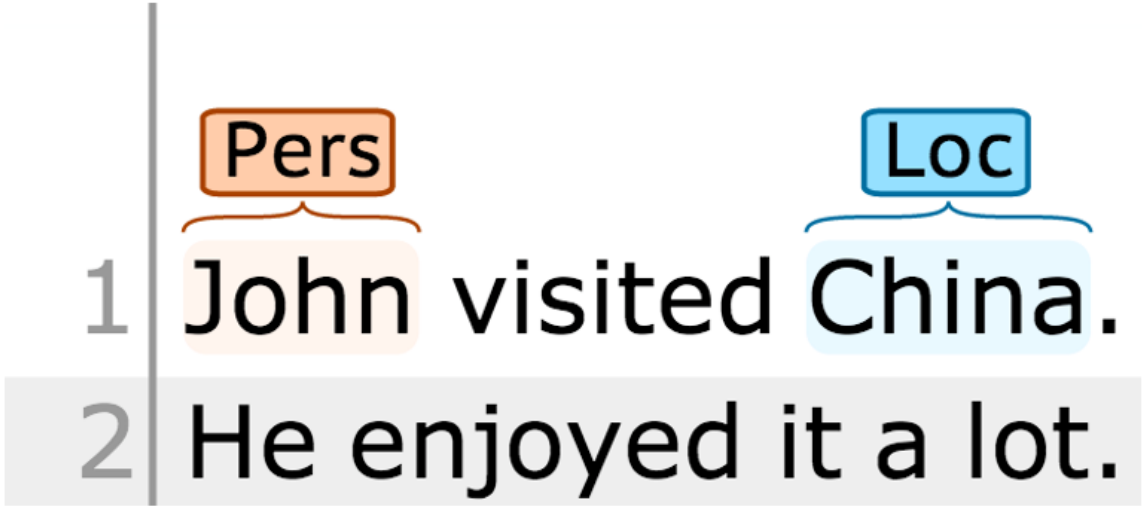
“John visited China. He enjoyed it a lot. ”

John visit China . He enjoy it a lot .

A run-through example

After named entity recognition

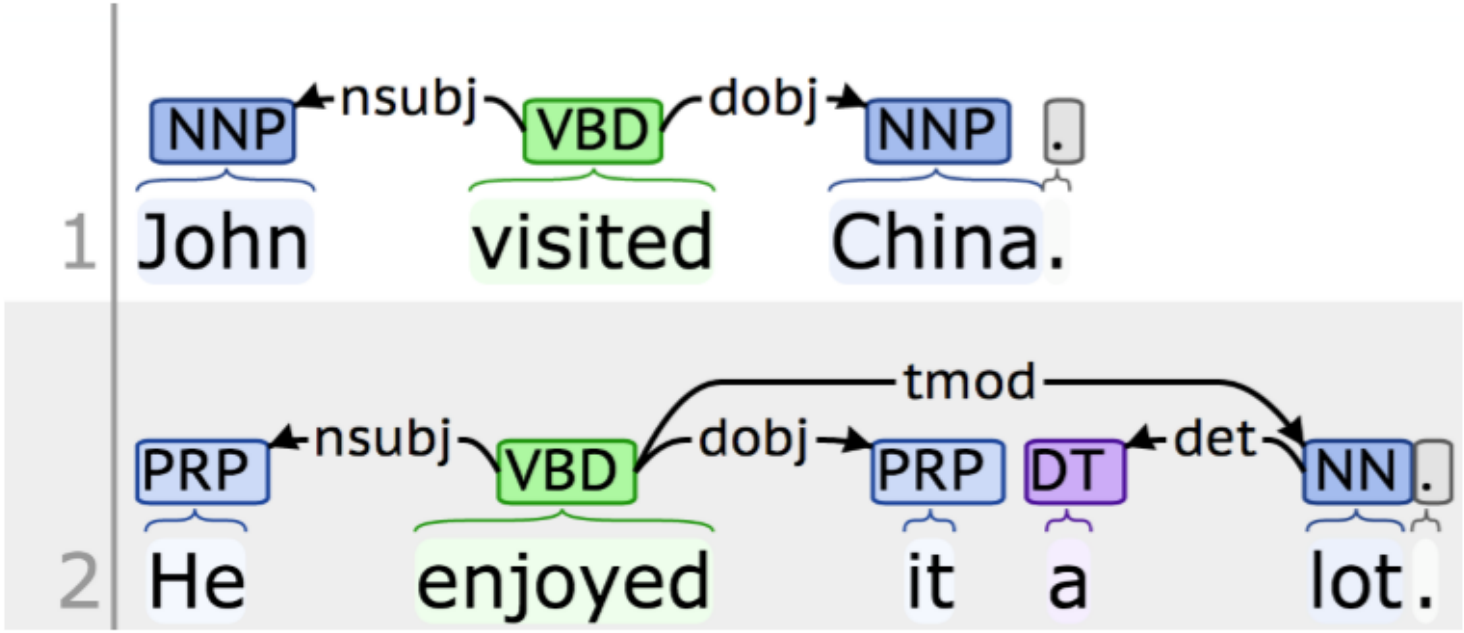
“John visited China. He enjoyed it a lot.”



A run-through example

After dependency parsing

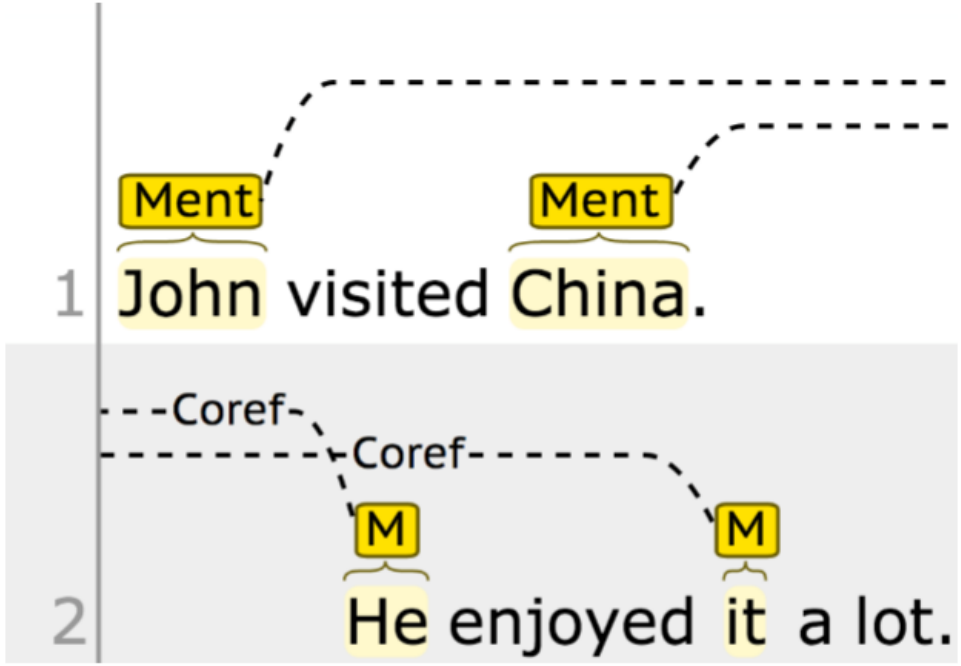
"John visited China. He enjoyed it a lot."



A run-through example

After coreference resolution

"John visited China. He enjoyed it a lot. "



Text processing packages

- NLTK (python)
  - http://nltk.org/
  - Has everything except coreference resolution
  - May not have state-of-the-art implementation

- Stanford's CoreNLP
  - http://nlp.Stanford.edu/software/corenlp.shtml
  - Includes everything
  - State-of-the-art implementations
  - Has python and R wrappers
  - Some components may not be the fastest or most memory efficient

Text processing packages

- OpenNLP
  - http://opennlp.apache.org
  - Includes everything
  - Very stable, but older algorithms
  - R package document: https://cran.r-project.org/web/packages/openNLP/openNLP.pdf

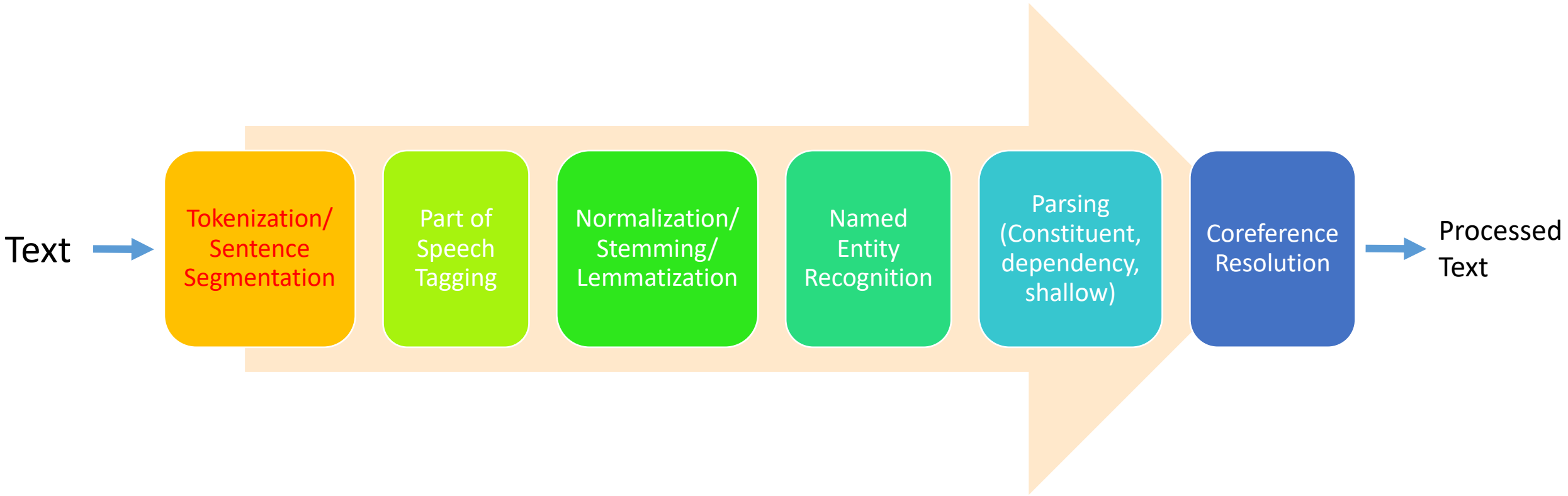
- TweetNLP
  - http://www.ark.cs.cmu.edu/TweetNLP
  - Best tools for working with tweets
  - Only tokenization and POS tagging

Talk with your neighbors

- Discuss with your neighbors for 5 min

- Do you have experience with text mining? What problems did you solve with text mining techniques?
- In your opinion, how do tech companies (e.g. Google) create value/revenue using text mining?
  - What are the data of interest?
  - What techniques do they use (and to do what)?

Typical text processing pipeline



Tokenization

- Segment running text into words and sentences
- First impulse: “Just segment around space characters.” Not so easy!

Mr. Sherwood said reaction to Sea Container’s proposal has been “very positive.” In New York Stock Exchange composite trading yesterday, Sea Containers closed at $62.625, up 62.5 cents.

- Segmenting on white spaces produces tokens such as: “cents.”, “said,”, “positive.””

Tokenization issues

- Many weird words
  - C++, C#, M*A*S*H
  - Emoticons: ☺ ;-)
- Contractions
  - I'll, isn't, dog's, etc
  - You want to split these in separate words because they impact grammar.
    - If not split, where's the verb in "I'm right."?

Tokenization issues

- Hyphenation
  - Some are clearly single words
    - e-mails, co-operate, A-1-plus
  - Arguable cases, usually single words
    - non-lawyer, pro-Arab, so-called
  - Hyphenation often used to indicate the correct grouping of words
    - the once-quiet study of superconductivity
    - a final “take-it-or-leave-it” offer
    - the 26-year-old

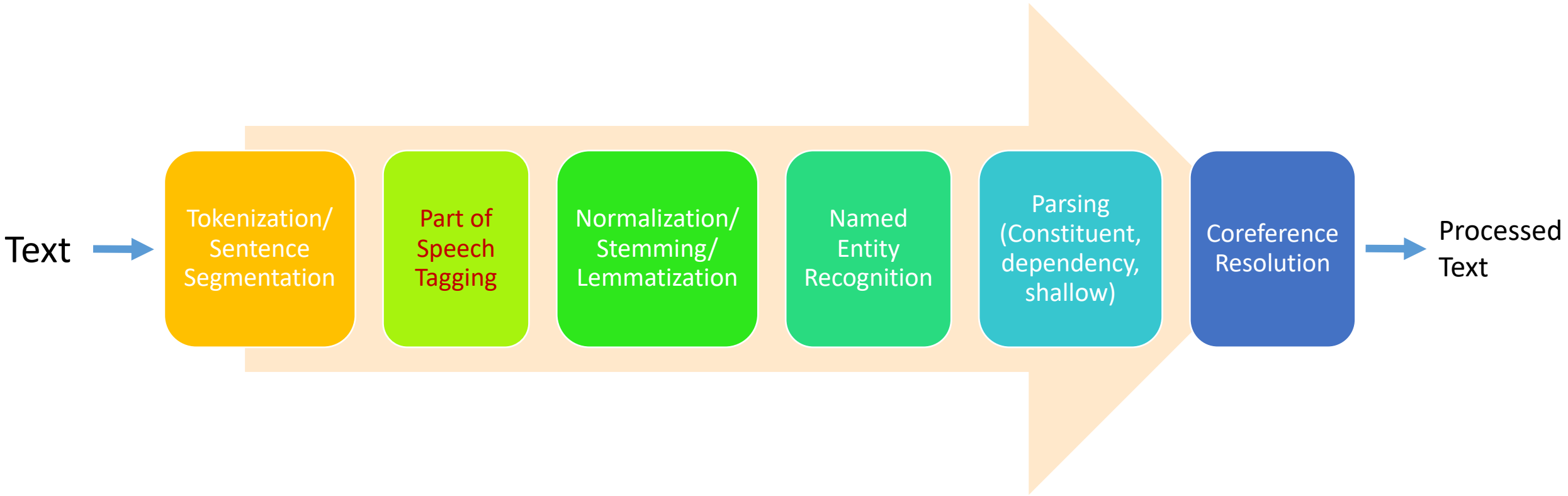
Tokenization issues

- Whitespace not indicating a word break
  - Multi-park names
    - “New York” is different from “York”
  - Phrasal verbs
    - Make up, work out
  - Phone numbers
    - USA: (202) 555-2230
    - France: 33 1 344 43 32 26
    - Sri Lanka: (94-1) 866854

NLTK Example

```
>>> sentence = '''At eight o'clock on Thursday morning Arthur didn't feel very g
ood.'''
>>> nltk.word_tokenize(sentence)
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning', 'Arthur', 'did', "n't",
'feel', 'very', 'good', '.']
```

Typical text processing pipeline



Part of Speech Tagging

- Process of performing automatic grammatical tagging for word categories

- *Part-of-speech tags* aka *word classes*, *morphological classes*, or *lexical tags*

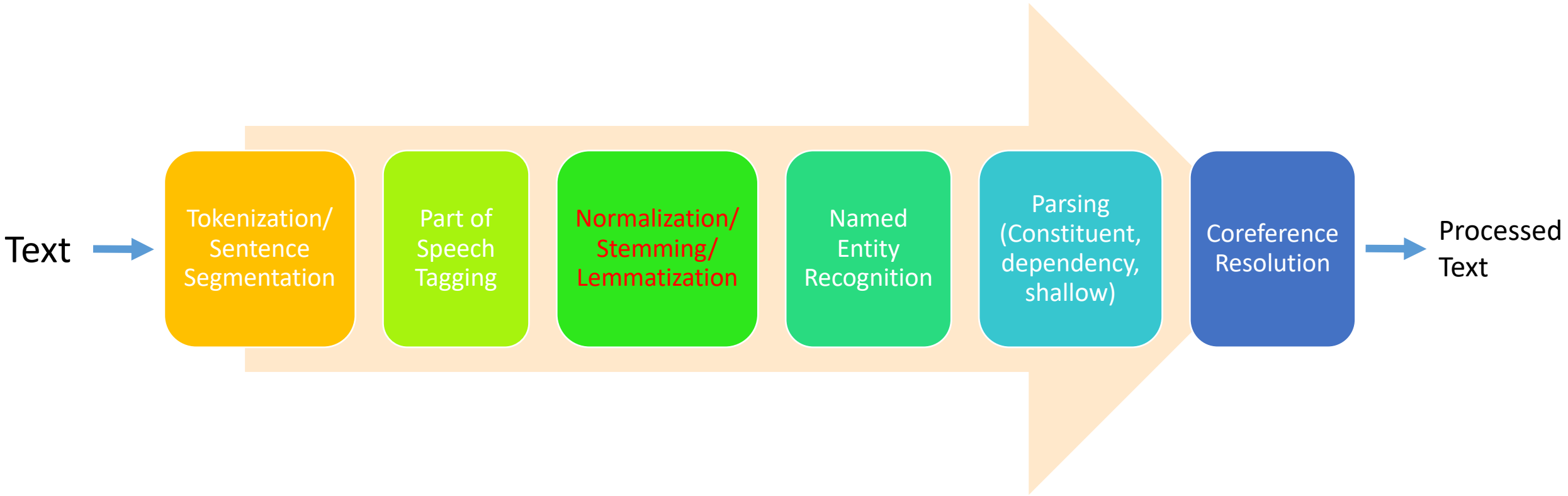
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

NLTK Example

Sentence = “At eight o’clock in Thursday morning Arthur didn’t feel very good.”

```
>>> token = nltk.word_tokenize(sentence)
>>> nltk.pos_tag(token)
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'NN'), ('on', 'IN'), ('Thursday', 'N
NP'), ('morning', 'NN'), ('Arthur', 'NNP'), ('did', 'VBD'), ("n't", 'RB'), ('fee
l', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')]
```

Typical text processing pipeline

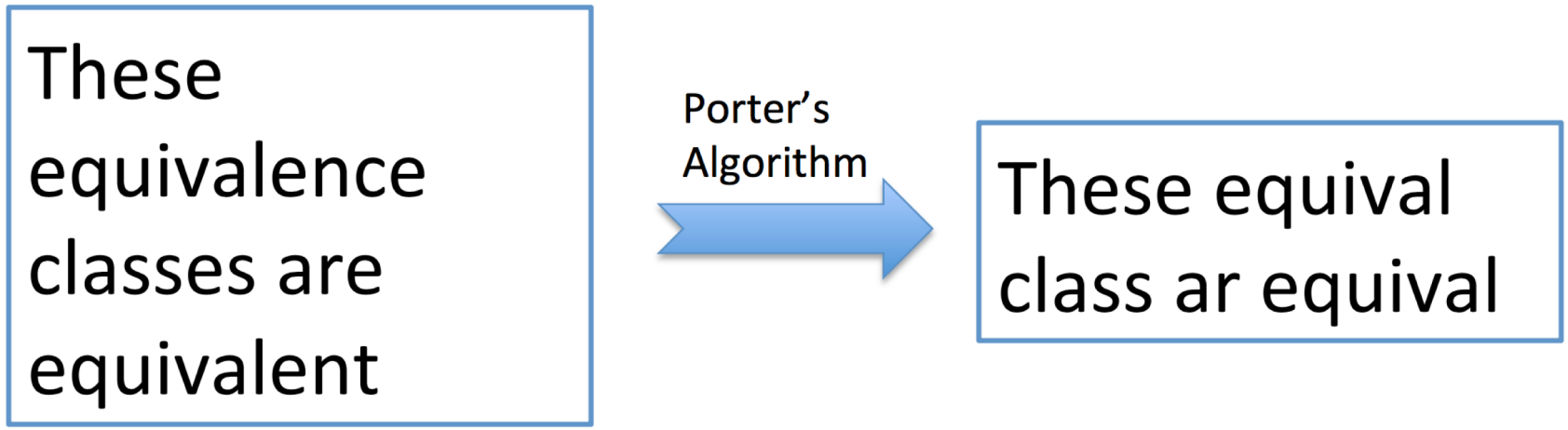


Normalization

- When searching a document, a query containing “USA” should match documents containing “U.S.A.”
- Implicit equivalence classes
  - Removing dots: “U.S.A.” ”USA”
  - Remove hyphens: “anti-discriminatory” “antidiscriminatory”
- Explicit equivalence classes
  - Different spellings: “color” and “colour”
  - Synonyms: “car” and “automobile”
- Case folding
  - USA -> usa

Stemming

- Reduce words to a common base form



Porter's Stemming Algorithm

- 5 phases of word reductions applied sequentially
  - sses → ss caresses → caress
  - ies → i ponies → poni
  - s → cats → cat

- Rules sensitive to the measure of a word
  - (m > 1) ement → replacement → replac
  - Does not change cement!

Lemmatization

- Grouping together the different inflected forms of a word so they can be analyzed as a single item.
- “walk”, “walked”, “walks”, “walking”
- lemma: “walk”

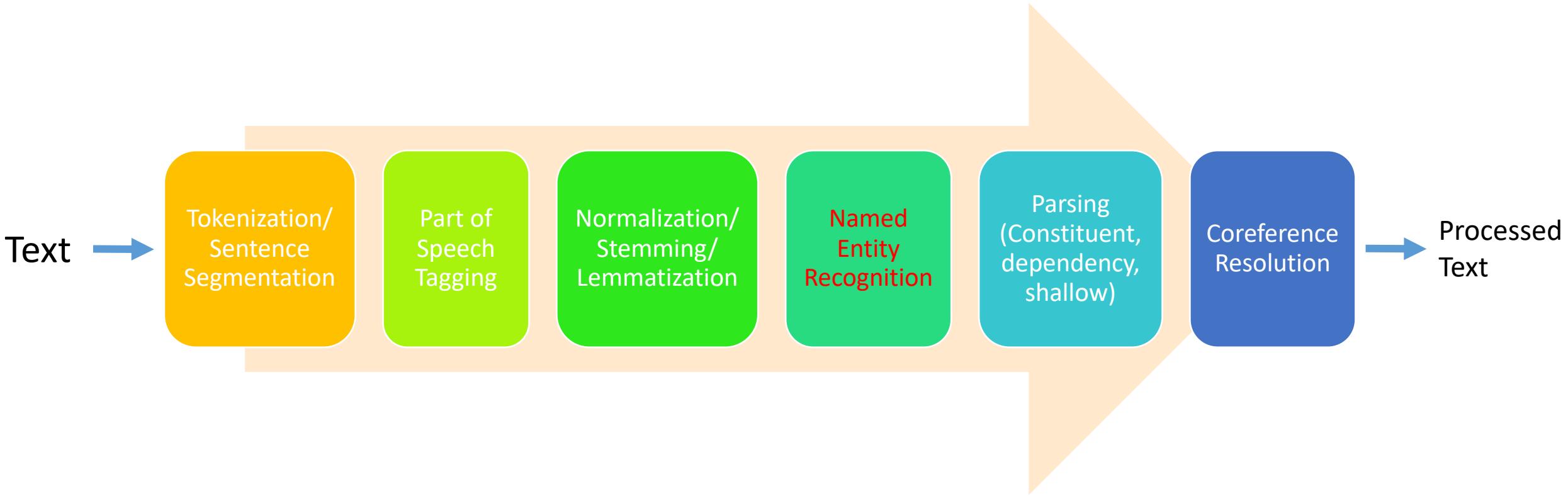
NLTK Example

Sentence = “At eight o’clock in Thursday morning Arthur didn’t feel very good.”

```
>>> from nltk.stem import PorterStemmer
>>> ps = PorterStemmer()
>>> for word in token:
...     print ps.stem(word)
...
At
eight
o'clock
on
thursday
morn
arthur
did
n't
feel
veri
good
.
```

```
>>> from nltk.stem.wordnet import WordNetLemmatizer
>>> lemmatizer = WordNetLemmatizer()
>>> for word in token:
...     print lemmatizer.lemmatize(word)
...
At
eight
o'clock
on
Thursday
morning
Arthur
did
n't
feel
very
good
.
```

Typical text processing pipeline



Named entity recognition

- Detection and classification of named entities in text

- In reality, a good NER will identify:
  - Named entities
  - Numeric entities
  - Temporal expressions

Named entity types

Type	Tag	Sample categories
People	PER	Individual, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-political entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicle	VEH	Planes, trains, automobiles

Named entity examples

Type	Example
People	**Turing** is often considered the father of computer science.
Organization	The **IPCC** said it is likely that future cyclones will be more intense.
Location	The **Mt. Sanitas** loop hike begins at the base of **Sunshine Canyon**.
Geo-political entity	**Palo Alto** will raise parking fees.
Facility	Drivers were advised to consider the **Lincoln Tunnel**.
Vehicle	The updated **Mini Cooper** retains its charm and agility.

Numeric entities

Type	Tag	Example
Money	MONEY	This laptop costs **$450**.
Number	NUMBER	He was the **42nd** president.

Temporal expressions

Type	Tag	Example
Time	TIME	This class starts at **8:30am**.
Date	DATE	The first lecture is on **August 26, 2013**.

NER as sequence labeling

- Most entity mentions span multiple sentences

- Dedicated representations
  - Most common: IOB
  - B: ‘beginning’
  - I: ‘inside’
  - O: ‘outside’

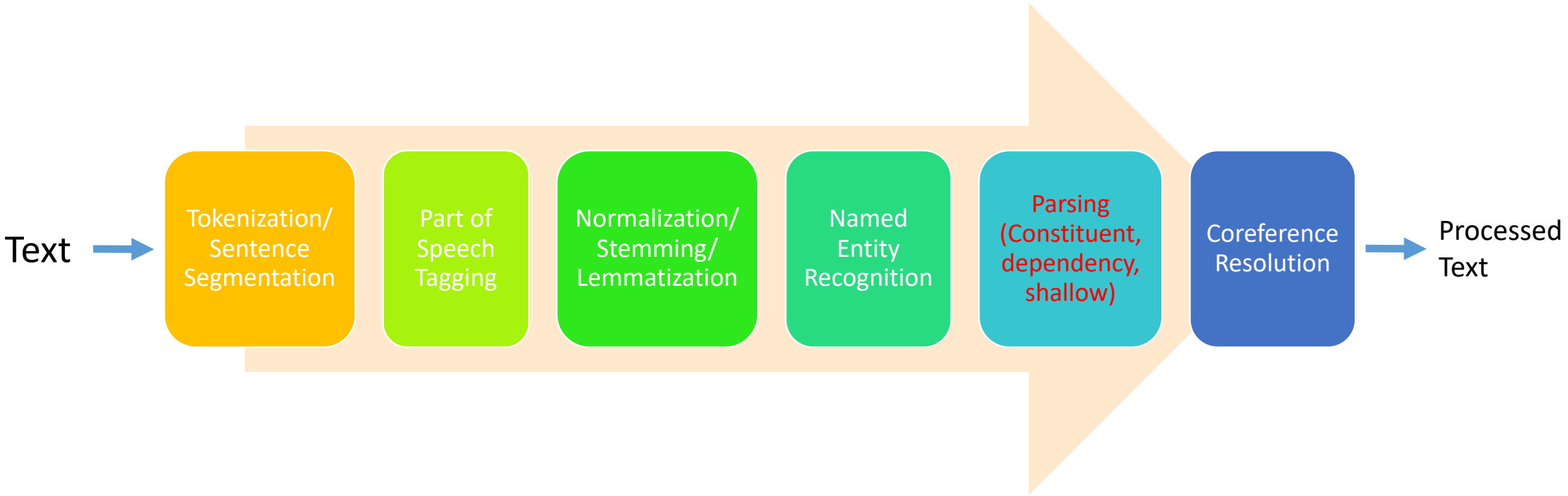
Words	Label
American	B-ORG
Airlines	I-ORG
,	O
a	O
unit	O
of	O
AMR	B-ORG
Corp.	I-ORG
,	O
immediately	O
matched	O
the	O
move	O
,	O
Tim	B-PER
Wagner	I-PER
said	O
.	O

NLTK Example

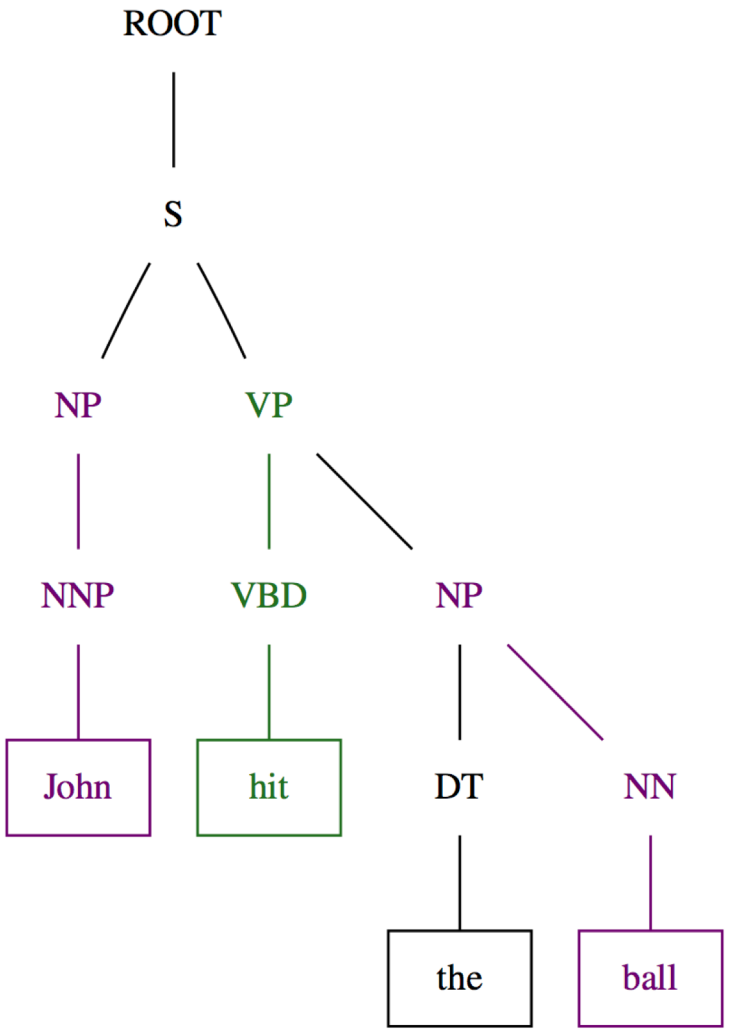
Sentence = “At eight o’clock in Thursday morning Arthur didn’t feel very good.”

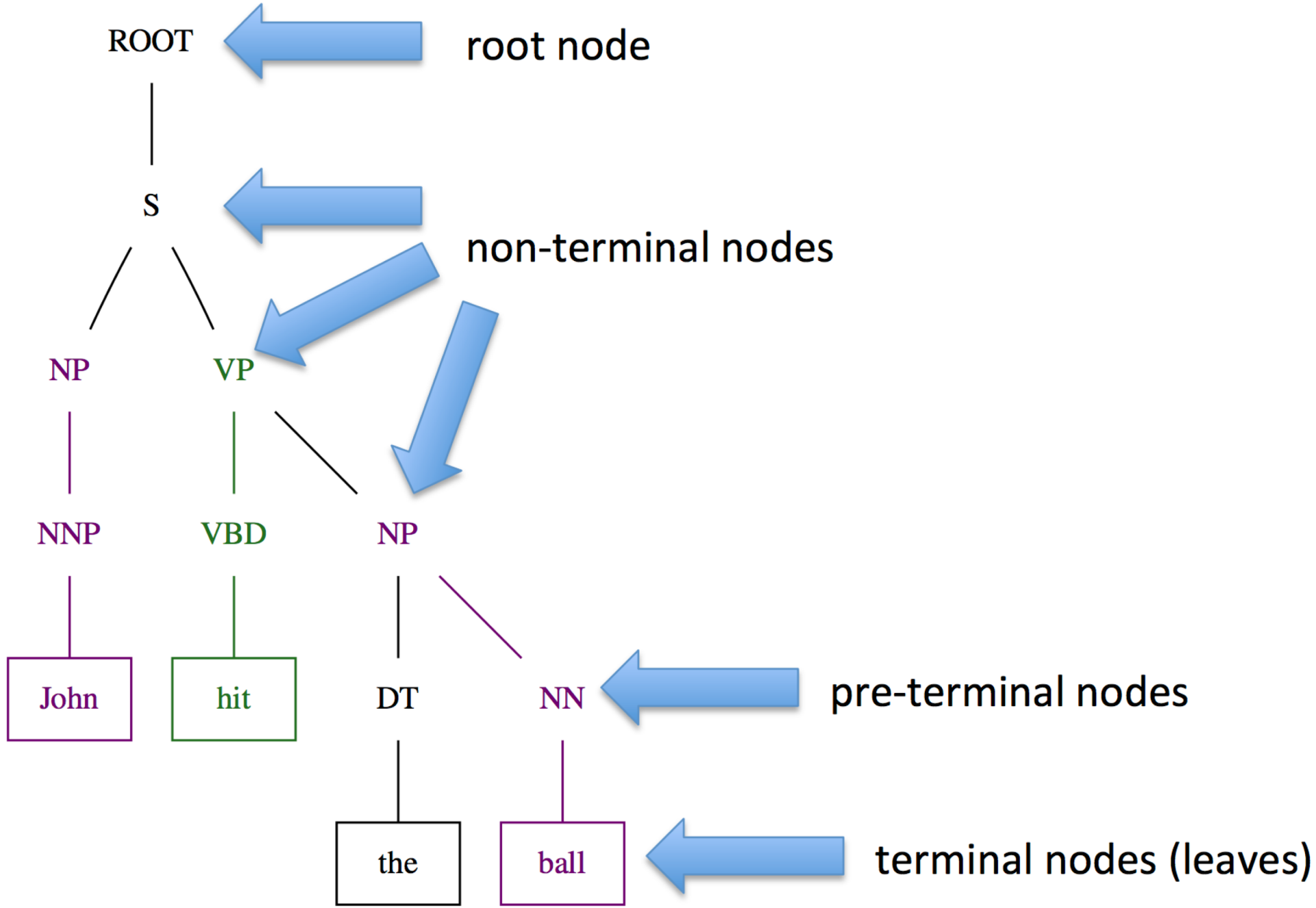
```
>>> from nltk import word_tokenize, pos_tag, ne_chunk
>>> print ne_chunk(pos_tag(word_tokenize(sentence)))
(S
  At/IN
  eight/CD
  o'clock/NN
  on/IN
  Thursday/NNP
  morning/NN
  (PERSON Arthur/NNP)
  did/VBD
  n't/RB
  feel/VB
  very/RB
  good/JJ
  ./.)
```

Typical text processing pipeline



Constituency-based parse tree

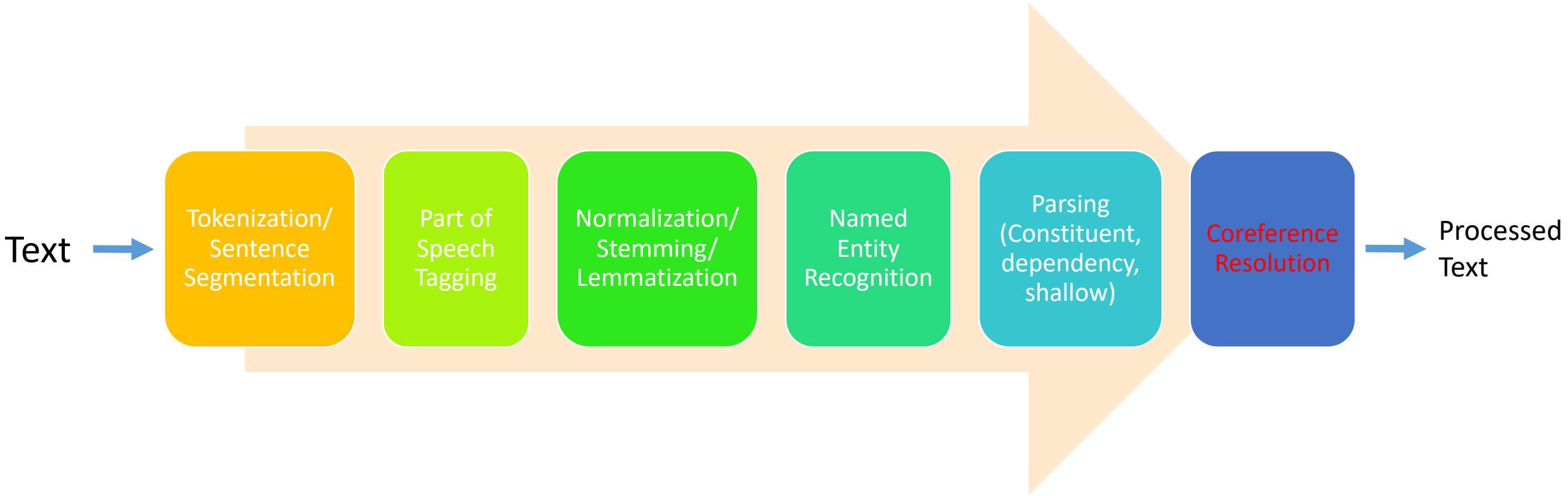



ROOT
root node
S
non-terminal nodes
NP
VP
NNP
VBD
NP
John
hit
DT
NN
pre-terminal nodes
the
ball
terminal nodes (leaves)

Applications of parsing

- Grammar checking (e.g., Grammarly)
  - I want to return **this** shoes
- Question answering
  - How many people in sales make $40k or more per year?
  - Need a parser to recognition that you want the sales, 40k is an attribute
- Machine translation (e.g, word order)
- Information extraction, speech generation, speech understanding, interpretation…

Typical text processing pipeline



Coreference resolution

Victoria Chen, Chief Financial Officer of MegaBucks Banking Corp. since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old became the Denver-based financial services company's president. It has been years since she came to Megabucks from rival LotsaBucks.

Coreference resolution

Victoria Chen, Chief Financial Officer of MegaBucks Banking Corp. since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old became the Denver-based financial services company's president. It has been years since she came to Megabucks from rival LotsaBucks.

- Coreferring expressions: *Victoria Chen*, *Chief Financial Officer of MegaBucks Banking Corp. since 2004*, *her*, *the 37-year-old*, *the Denver-based financial services company's president*
- Referent: **Victoria Chen**

Coreference resolution

Victoria Chen, Chief Financial Officer of MegaBucks Banking Corp. since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old became the Denver-based financial services company's president. It has been years since she came to Megabucks from rival LotsaBucks.

- Coreferring expressions: *MegaBucks Banking Corp.*, *the Denver-based financial services company*, *MegaBucks*
- Referent: **MegaBucks Banking Corp.**

Definition

- Coreference resolution
  - Finding expressions that corefer, i.e., referring expressions that refer to the same entity

- An important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction

Summary

- Text processing
  - Tokenization
  - Part-of-speech tagging
  - Normalization/stemming/lemmatization
  - Named entity recognition
  - Parsing
  - Coreference resolution

- Text mining and Web mining techniques