part-of-speech tagging, dependency and shallow parsing

André Santos, afs@inesctec.pt

POS tagging

Part-of-speech tagging

(or just tagging for short) is the process of assigning a part-of-speech or other syntactic class marker to each work in a corpus.

Jurafsky & Martin, Speech and Language Processing

POS tagging

Part-of-speech tagging

(or just tagging for short) is the process of assigning a part-of-speech or other syntactic class marker to each work in a corpus.

Jurafsky & Martin, Speech and Language Processing

[text] corpus

(pl. corpora) is a large and structured set of texts

POS in Portuguese

- substantivo,
- artigo,
- adjetivo,
- numeral,
- pronome,
- verbo,
- advérbio,
- preposição,
- conjunção,
- interjeição

Tagsets for Portuguese

Freeling

- **A**: adjective
- **C**: conjunction
- **D**: determiner
- **N**: noun
- **P**: pronoun
- **R**: adverb
- **S**: adposition
- **V**: verb
- **Z**: number
- **W**: date
- **I**: interjection

Tagsets for Portuguese

- A lot more options inside each category
- E.g. **nouns**:

POSITION	ATRIBUTE	VALUES
0	category	N: noun
1	type	C: common; P: proper
2	gen	F: feminine; M: masculine; ...
3	num	S: singular; P: plural; ...
4	neclass	S: person; G: location; ...
5	nesubclass	Not used
6	degree	A: augmentative; D: diminutive

POS taggers

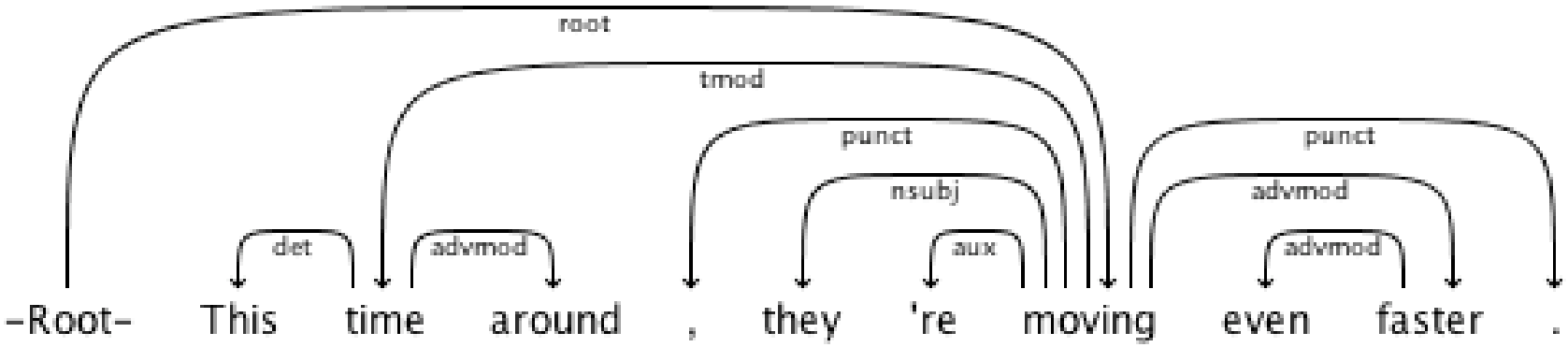
rule-based: taggers generally involve a large database of hand-written disambiguation rules

stochastic: taggers generally resolve tagging amiguities by using a training corpus to compute the probability of a given word having a given tag in a given context

transformation-based: (or Brill) tagger shares features from both previous approaches: it automatically induces rules from a previously tagged training corpus

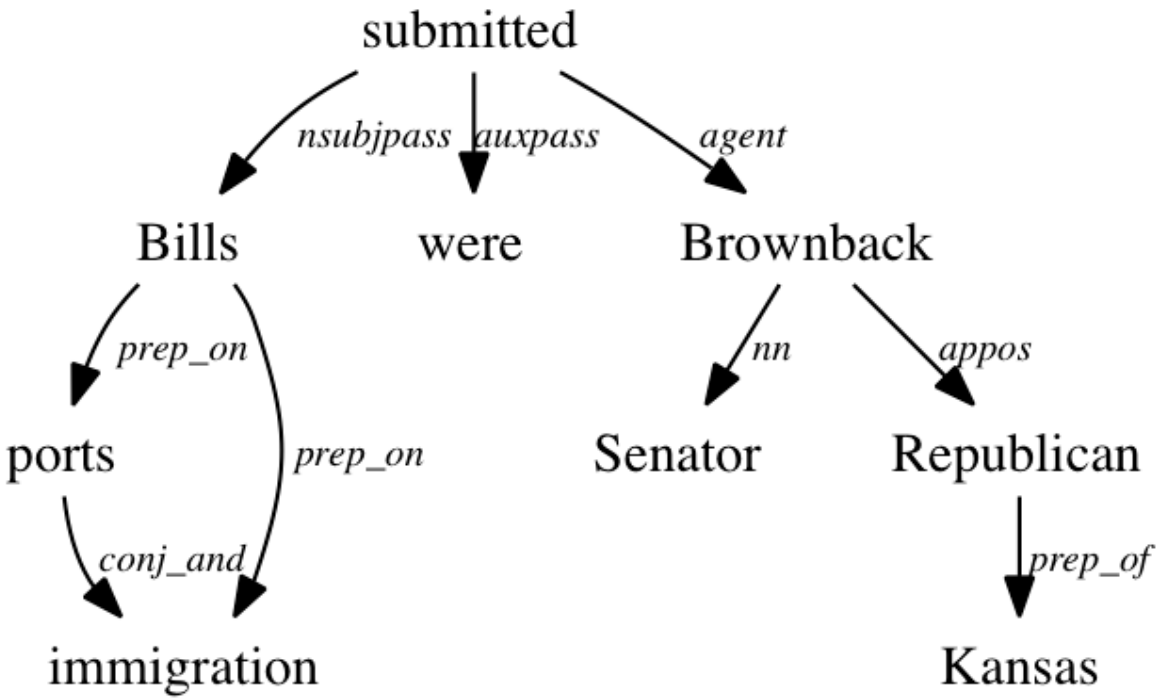
Dependency parsing

A dependency parser analyzes the grammatical structure of a sentence, establishing relationships between "head" words and words which modify those heads.



Dependency parsing

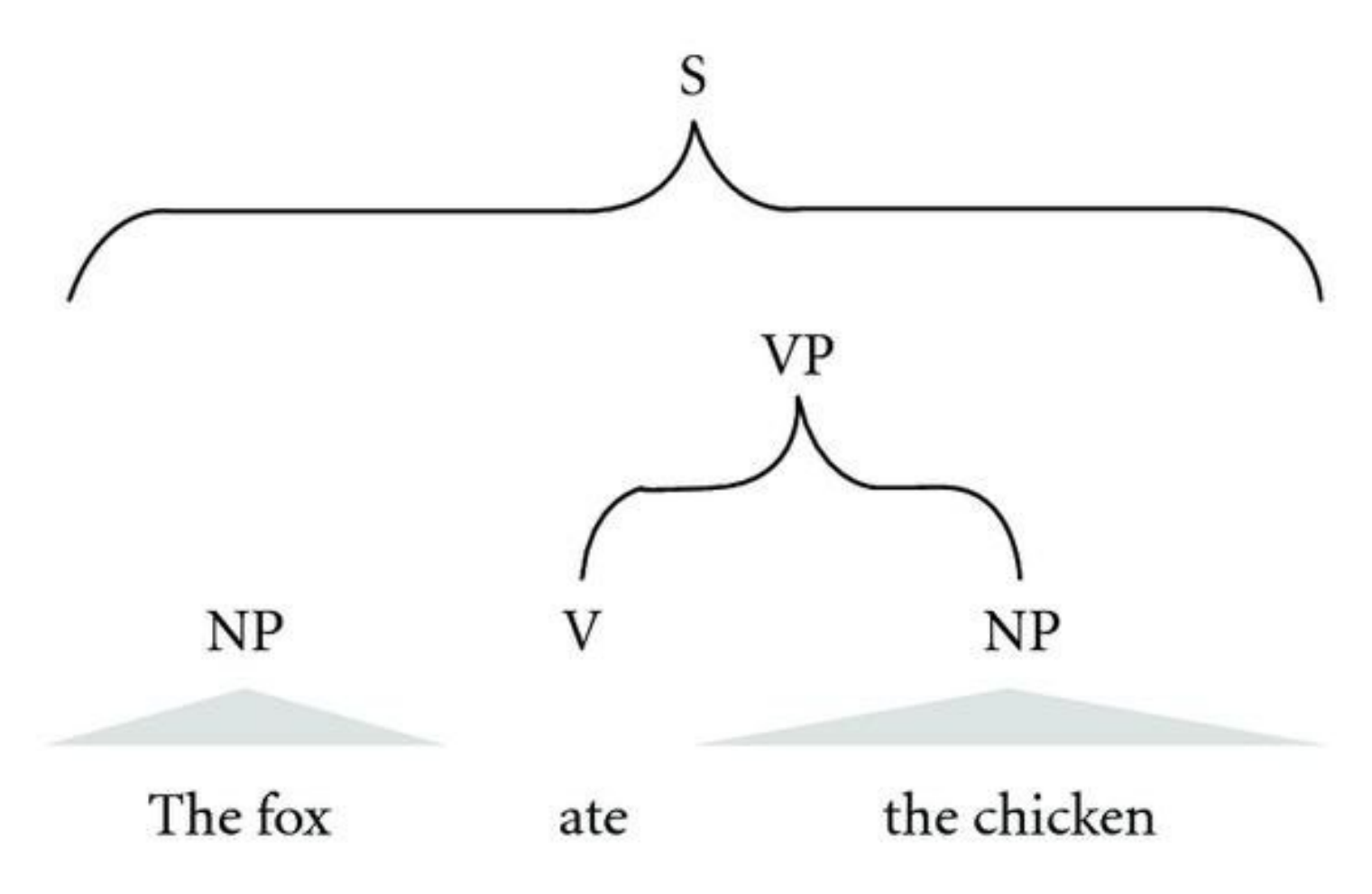
"Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."



Shallow parsing

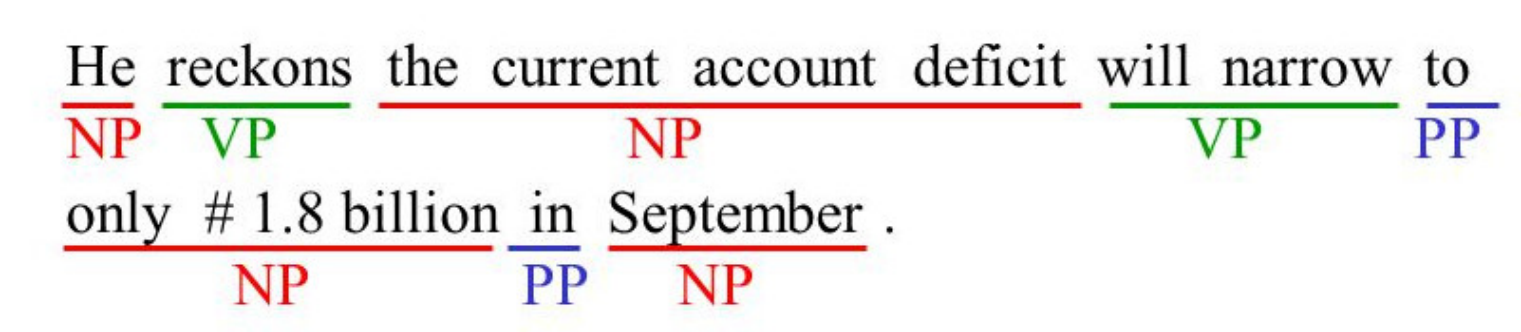
identifies the constituents (noun groups, verbs, verb groups, etc.), but does not specify their internal structure, nor their role in the sentence.

Shallow parsing



Chunking

- **identification and classification of the major POS elements**
- **segments a sentence into non-recursive phrases**



Practical assignment #2

Explore, test and present an NLP framework or a Python module

- **For simplicity, we will assume the groups will be the same as TP#1 (but you can switch groups)**
- **Each group will be randomly assigned a topic (e.g. Freeling NER, `beautifulsoup`, `nltk`'s POS tagger, ...)**
- **Don't like your topic? Find a group to switch topics with**

Practical assignment #2

- **Presentation + report**
  - **Describe and explain how the tool/module works, and provide an NLP working example**
- **Presentation + submission date: Nov 30th 2018**
- **More info will be at** `github.com/andrefs/spln-2018-i/tree/master/data/assignments/2`

Exercises

1. **Build a `nl_grep` tool, which tags text and `grep`'s by POS**
   - **1.1** `nl_grep '%NP %V %N'`
   - **1.2** `nl_grep '%NP é %DET? %A? %N'`
   - **1.3** …