Second Assignment

Chiara Solito, VR487795

1 Prompts

The assignment consists in the development, in NLTK, OpenNLP, SketchEngine or GATE/Annie a pipeline that, starting from a text in input, in a given language (English, French, German and Italian are admissible) outputs the syntactic tree of the sentence itself, intended as a tree with root in S for sentence, and leaves on the tokens labelled with a single Part-of-speech. The generation of the tree can pass through one of the following models:

1. **PURE SYMBOLIC** The tree is generated by a LR analysis with CF LL2 grammar as a base. Candidates can assume the following:

   (a) Adjectives in English and German shall be only prefixed to nouns, whilst in French and Italian are only suffixed;

   (b) Verbs are all at present tense;

   (c) No pronouns are admitted;

   (d) Only one adverb is admitted, always post-poned with respect to the verb (independently of the language, and the type of adverb);

   Overall the point above map a system that could be devised in regular expressions, but a Context-free grammar would be simpler to define. Candidate can either define a system by themselves or use a syntactic tree generation system that can be found on GitHub. Same happens for POS-tagging, where some of the above mentioned systems can be customized by existing techniques that are available in several fashions (including a pre-defined NLTK and OpenNLP libraries for POS-tagging and a module in GATE for the same purpose. Ambiguity should be blocked onto first admissible tree.

2. **PURE ML** Candidates can develop a PLM with one-step Markov chains to forecast the following token, and used to generate the forecast of the POS tags to be attributed. In this case the PLM can be generated starting with a Corpus, that could be obtained online, for instance by using the Wikipedia access API, or other available free repos (including those available with SketchEngine. In this approach, candidates should never use the forecasting to approach the determination of outcomes (for this would be identical purpose of distinguishing EN/non ENG (and then IT/non IT, FR/not FR or DE/not DE) but only to identify the POS model in a sequence. In this case, the candidate should output the most likely POS tagging, without associating the sequence to a tree in a direct fashion.

Candidates are free to employ PURE ML approach to simplify, or pre-process the text in order to improve the performance of a PURE SYMBOLIC approach while generating a mixed model.

2 Delivering of the code

The code is in a public repository on GitHub:
ChiaraSolito/NLPAssignments/Assignment2. In the repository there are two versions of the code:[1]

- A notebook version, with all the comments from this document directly embedded.
- A python script version, that prints on terminal a formatted version of results.

[1] https://github.com/ChiaraSolito/NLPAssignments/tree/main/Assignment2

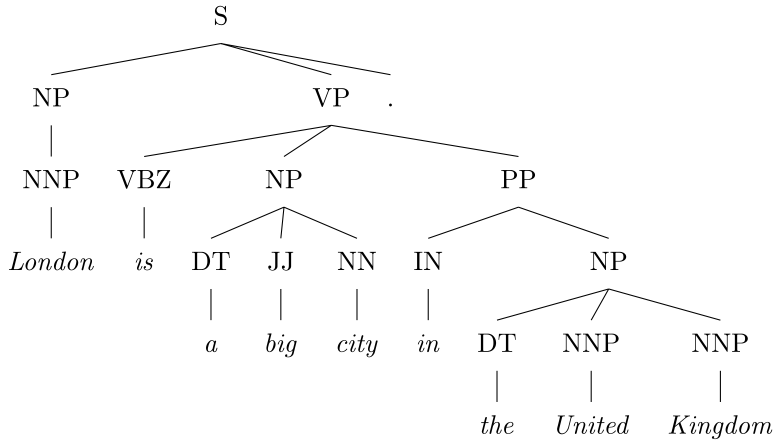
3 Approach

The approach chosen was PURE ML, using spaCy pipelines. The spaCy model is only used for tokenization and sentence segmentation, for parsing Berkeley Neural Parser is used. As specified in benepar documentation the reason is that parser models do not ship with a tokenizer or sentence splitter, and some models may not include a part-of-speech tagger either. After parsing, NLTK standard tree print is used, to print in the tree-format.

4 Results

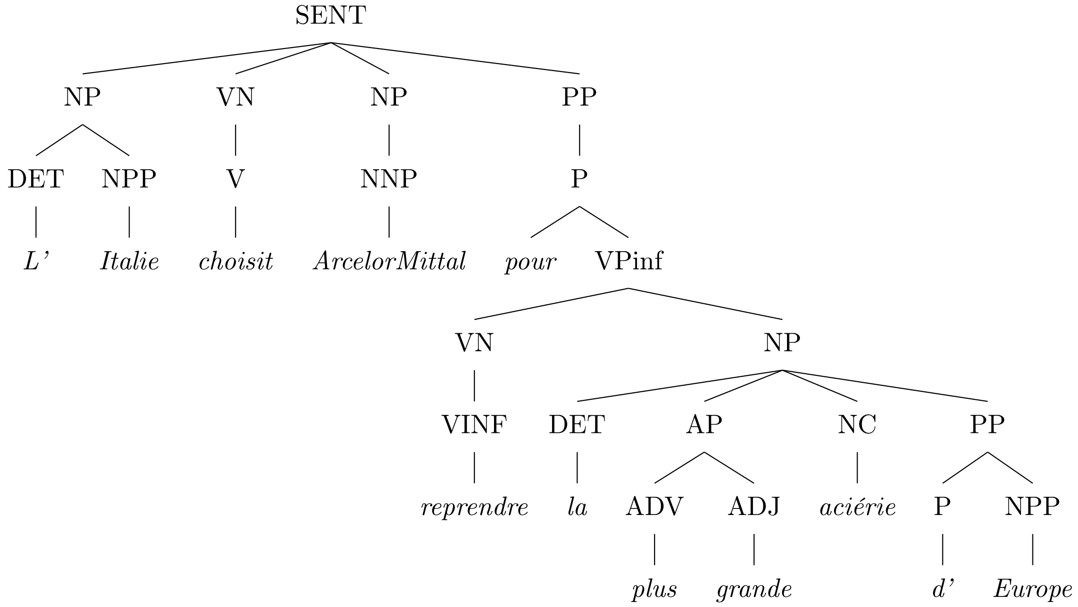
4.1 English

Sentence: ***London is a big city in the United Kingdom.***



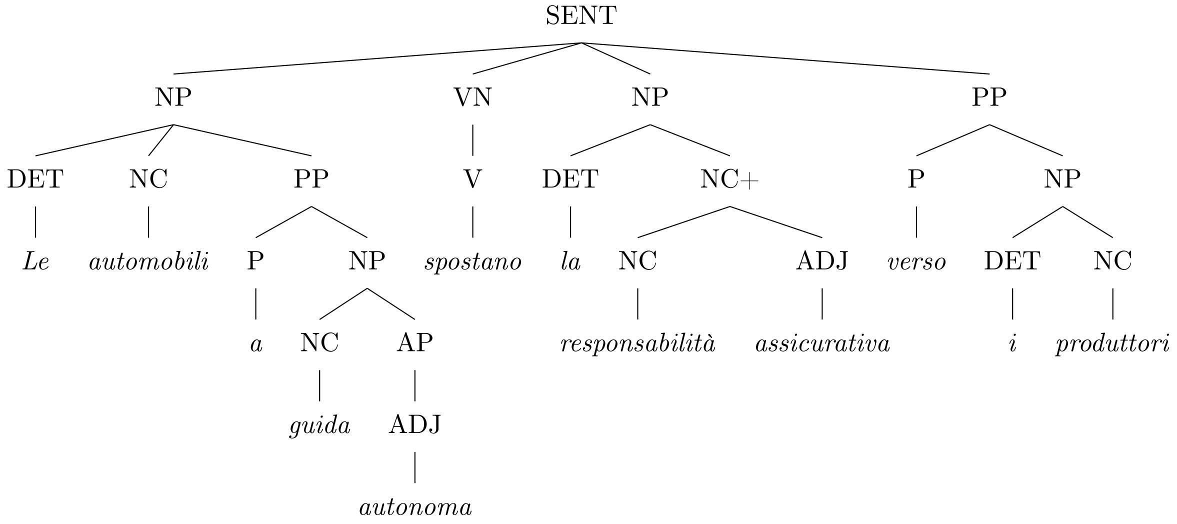
4.2 French

Sentence: ***L'Italie choisit ArcelorMittal pour reprendre la plus grande aciérie d'Europe***



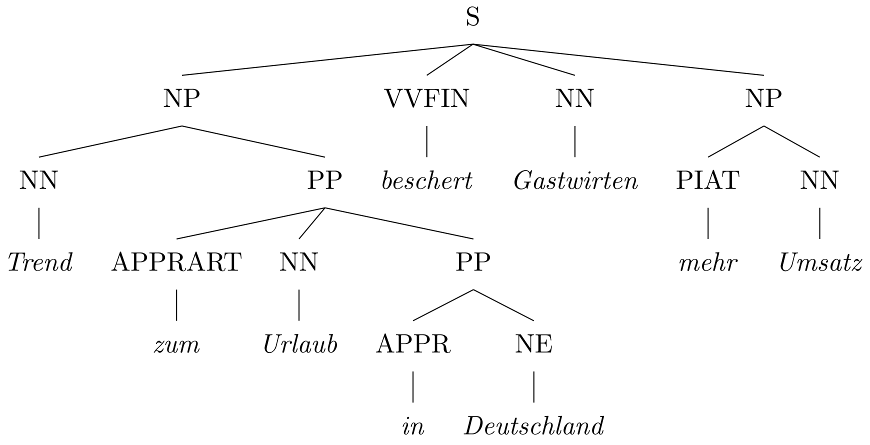
4.3 Italian

Sentence: ***Le automobili a guida autonoma spostano la responsabilità assicurativa verso i produttori***



4.4 German

Sentence: ***Trend zum Urlaub in Deutschland beschert Gastwirten mehr Umsatz***


S
NP
VVFIN
NN
NP
NN
PP
beschert
Gastwirten
PIAT
NN
Trend
APPRART
NN
PP
mehr
Umsatz
zum
Urlaub
APPR
NE
in
Deutschland