Semi-automated Extraction of a Wide-Coverage Type-Logical Grammar for French

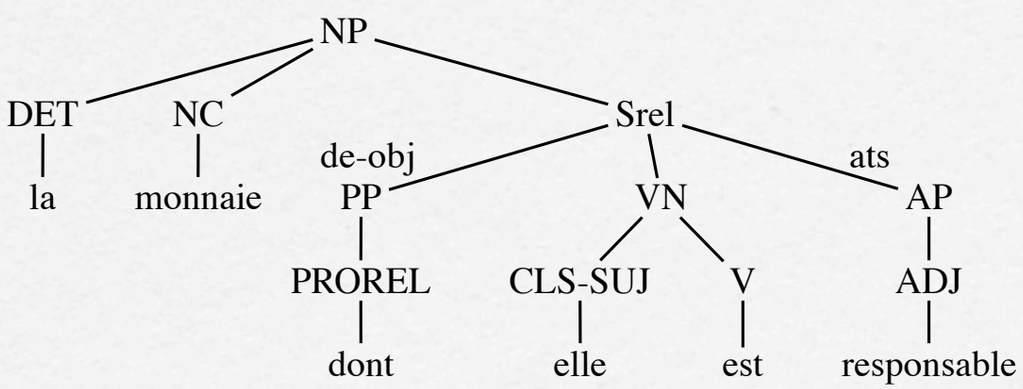
Richard Moot
LaBRI (CNRS, Bordeaux), SIGNES (INRIA Bordeaux SW)
Richard.Moot@labri.fr

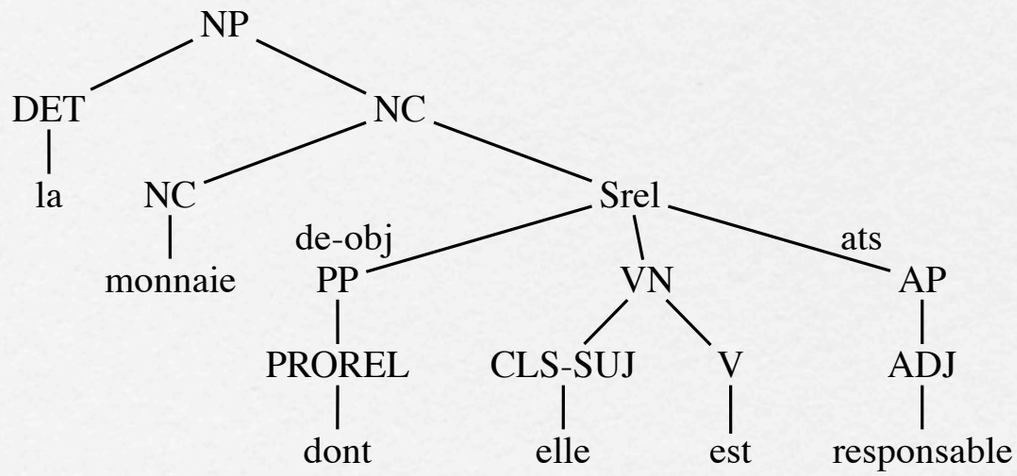
Introduction

- This "booster" describes a wide-coverage type-logical grammar for French, which is being developed with semantic applications in mind
- I will briefly describe the extraction algorithm, the extracted grammar and a first evaluation of the the grammar using a supertagger

The extraction algorithm

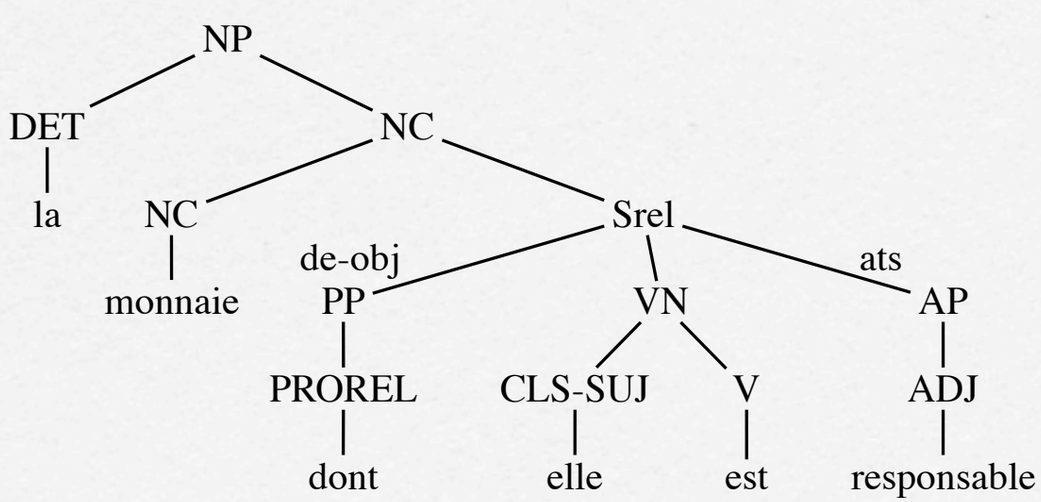
1. Binarize the annotation

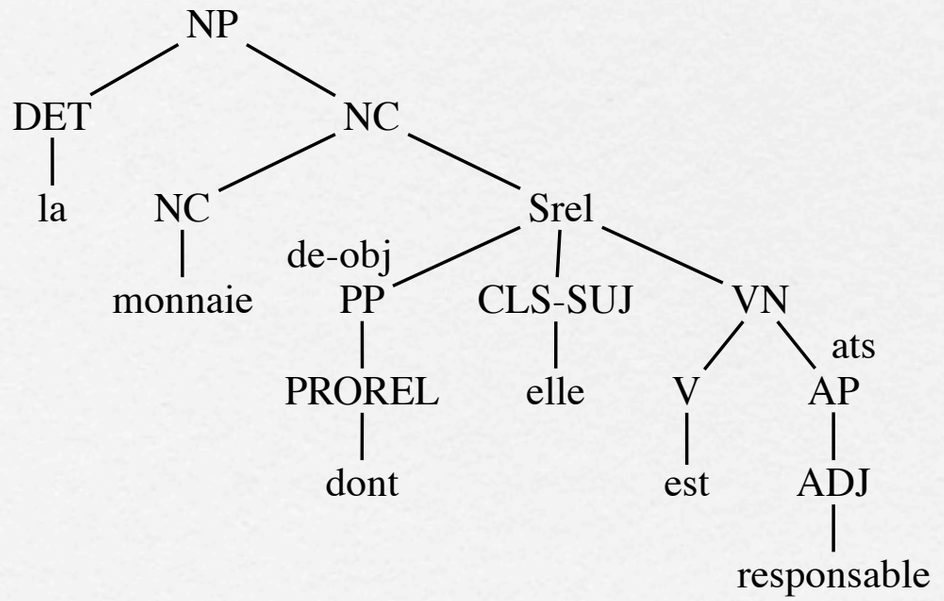




The extraction algorithm

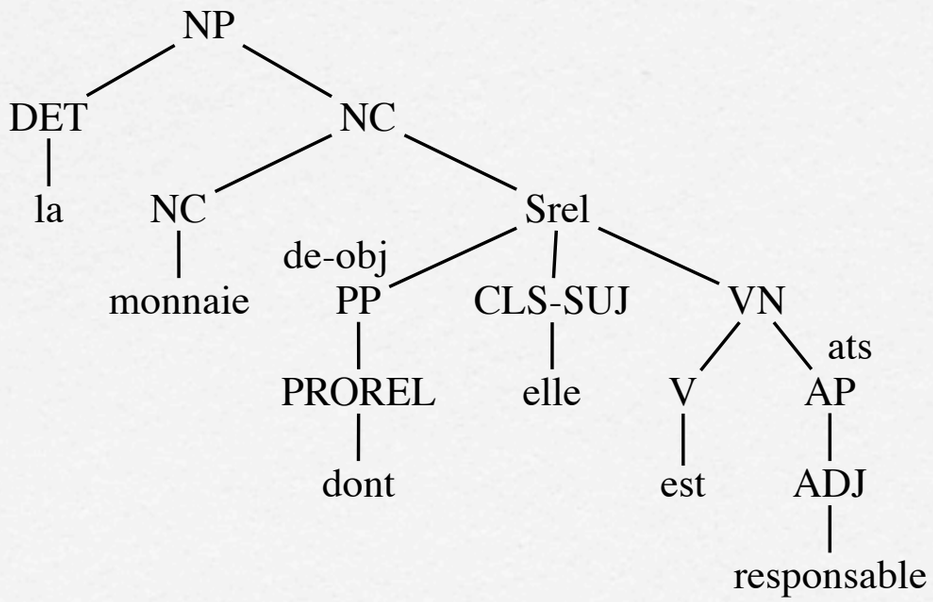
1. Binarize the annotation

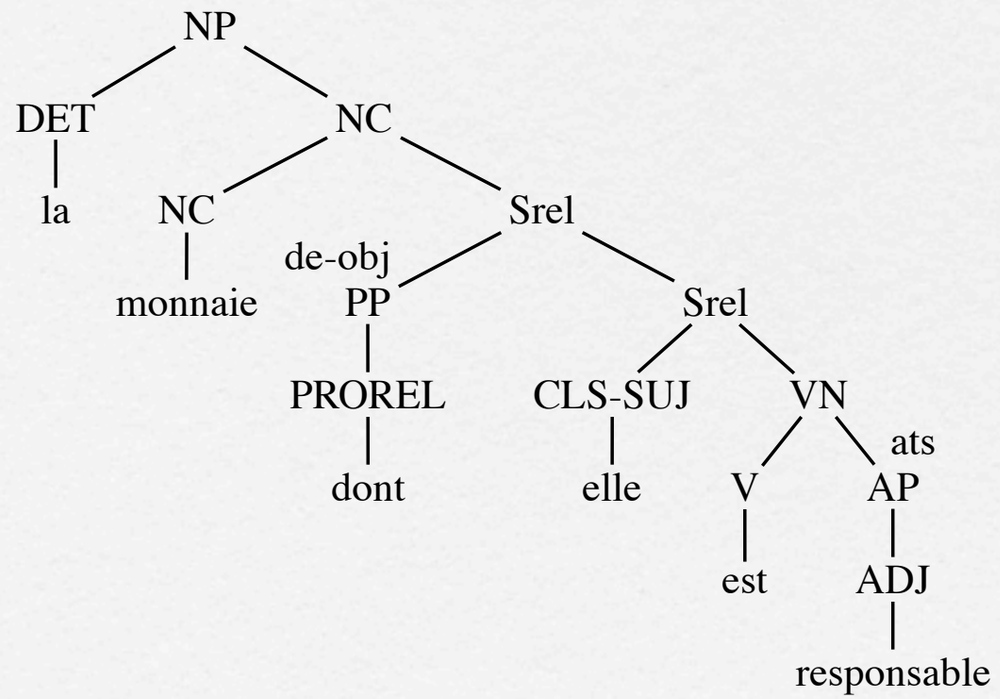




The extraction algorithm

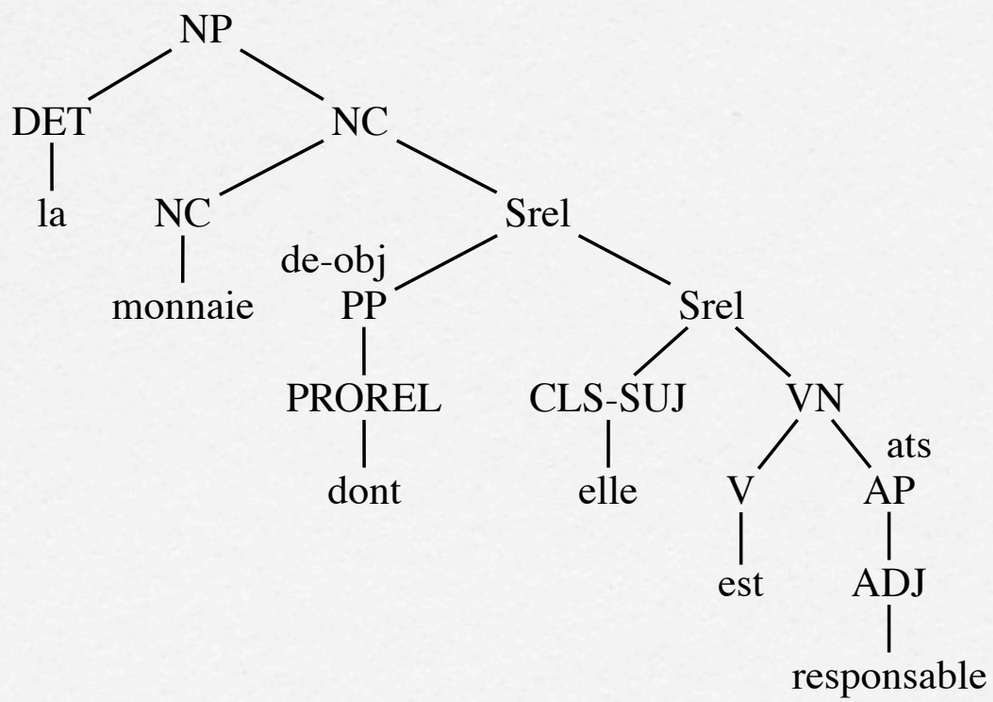
1. Binarize the annotation

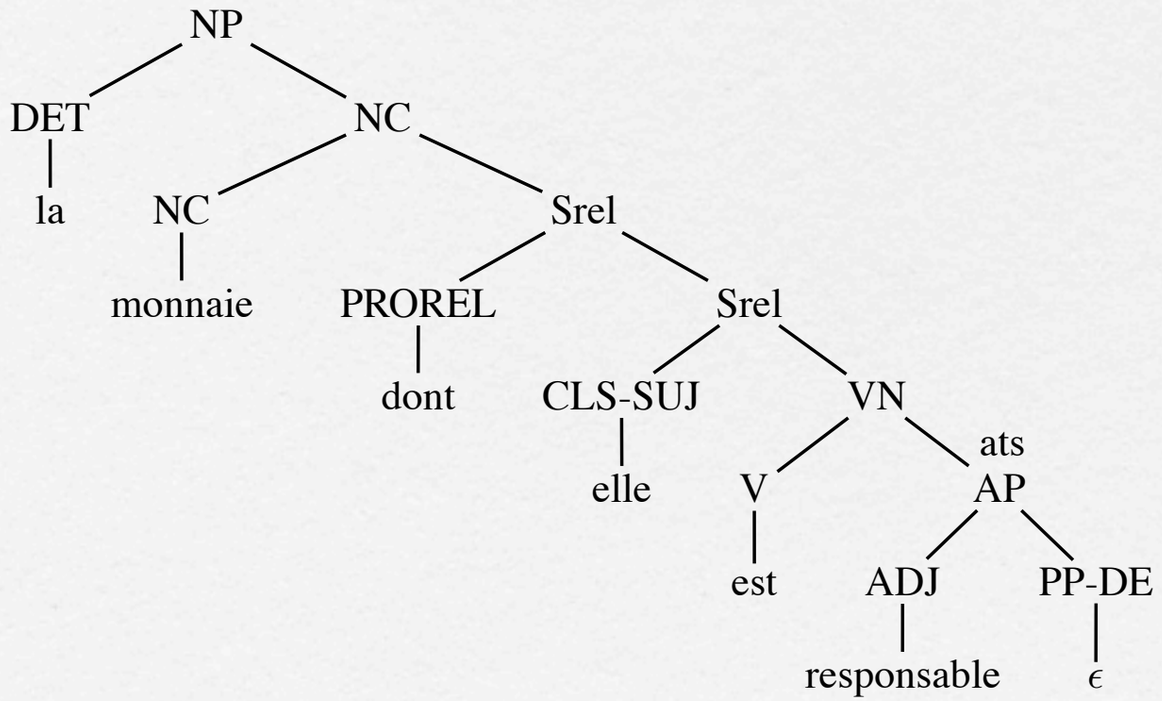




The extraction algorithm

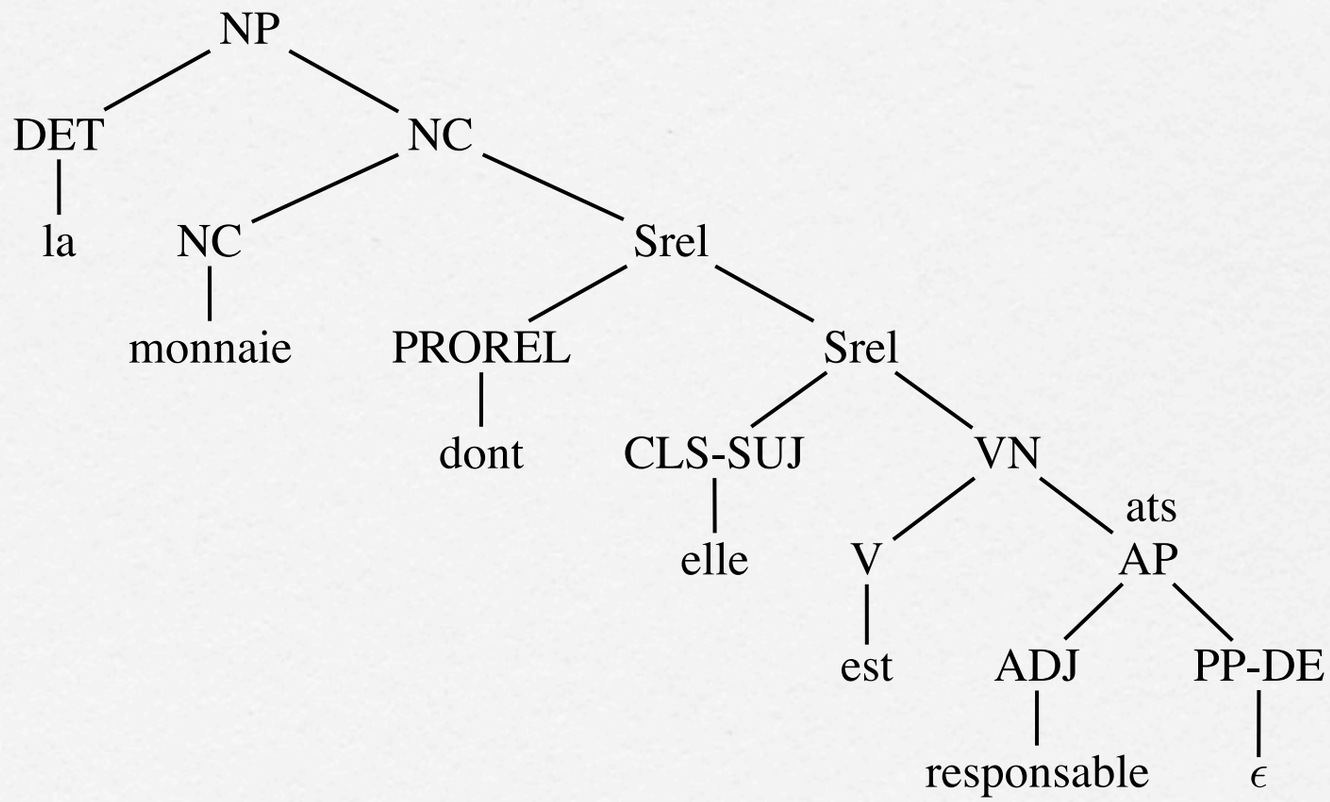
1. Binarize the annotation
   inserting "traces" for wh words





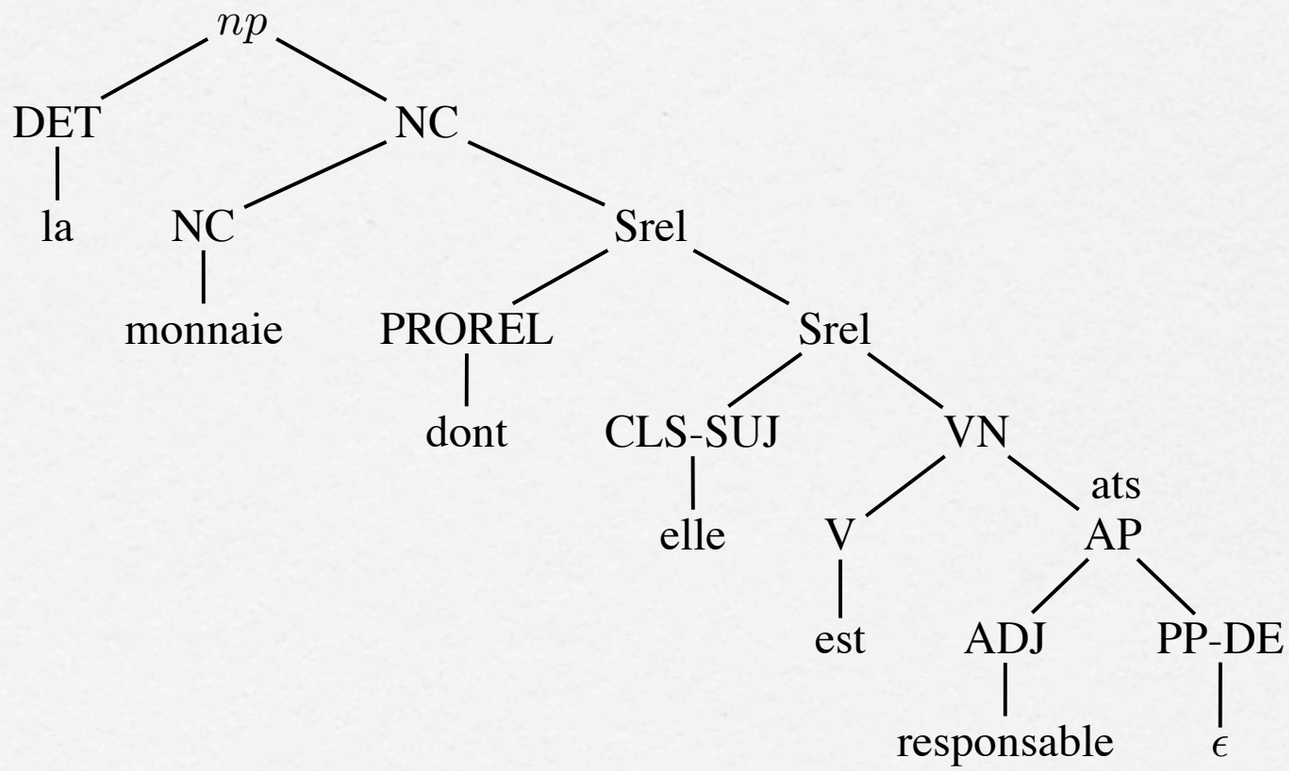
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



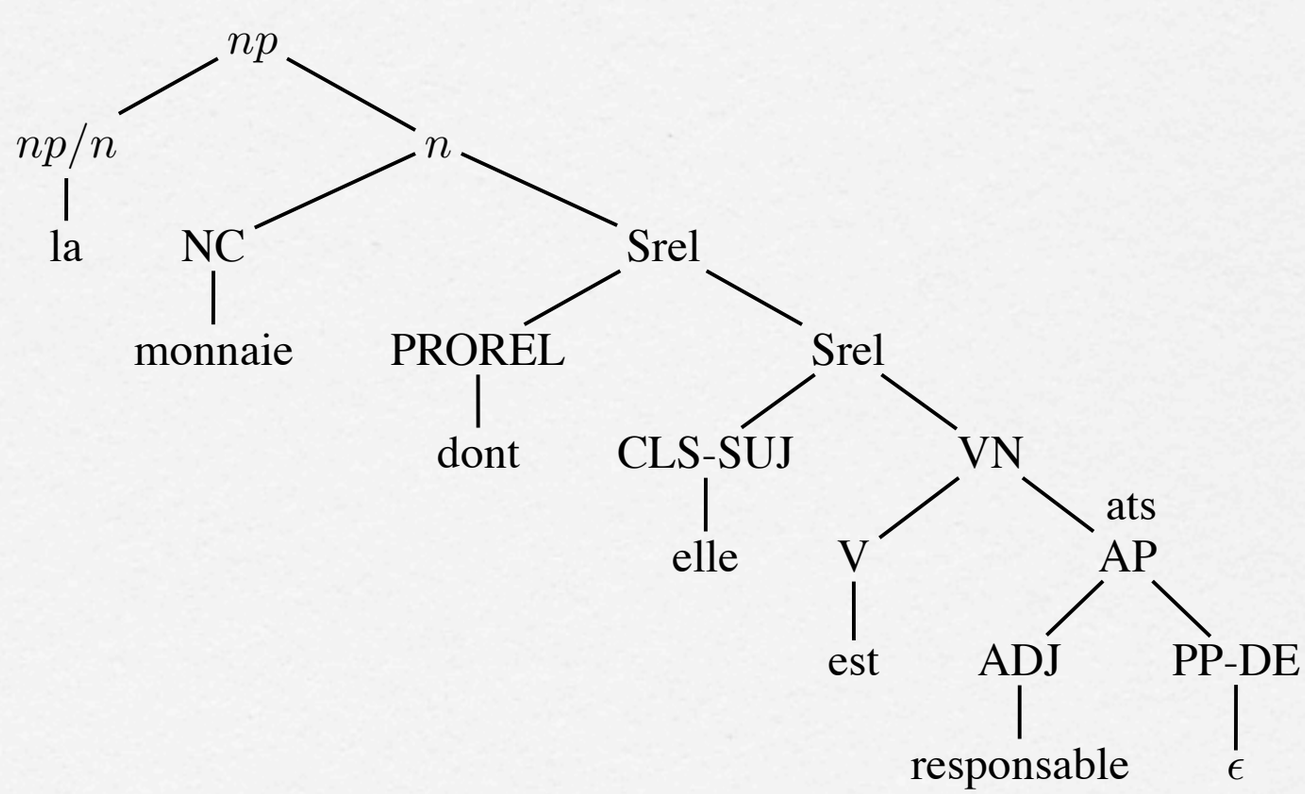
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



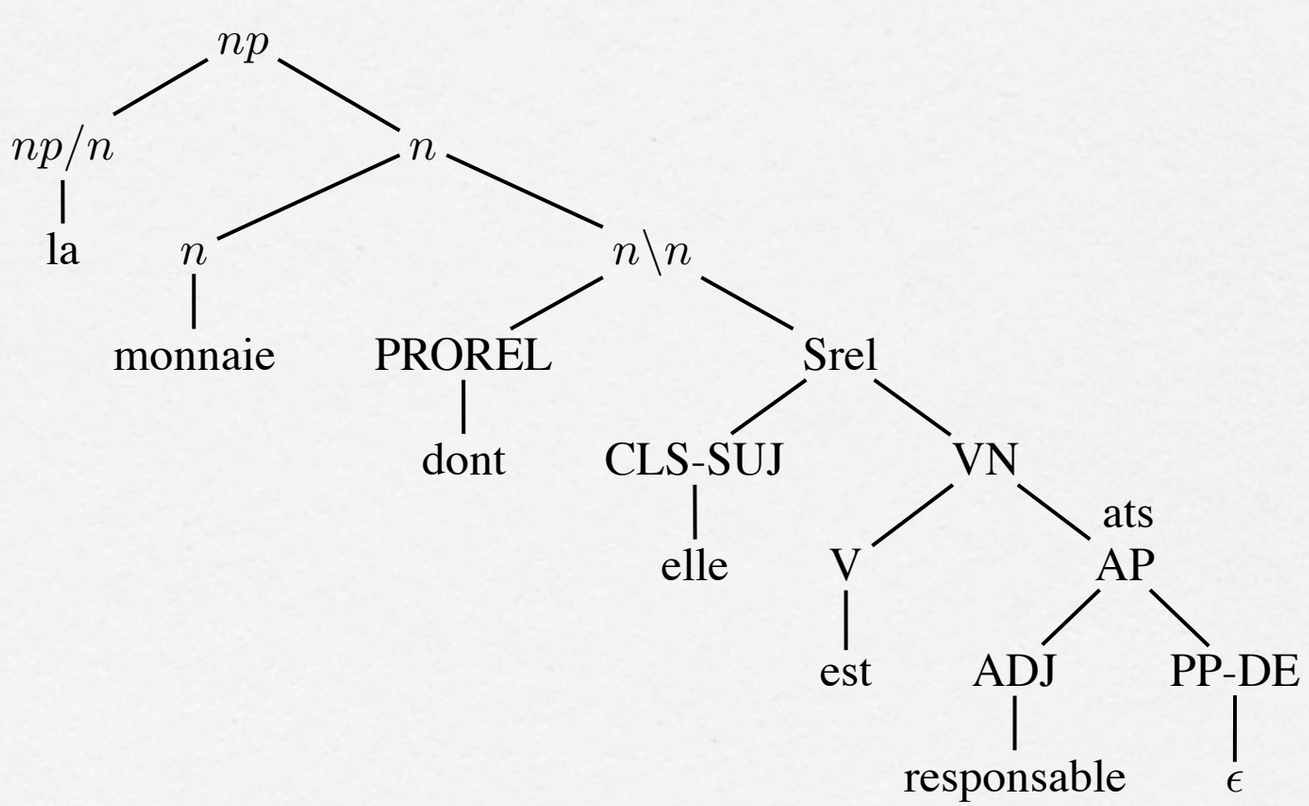
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



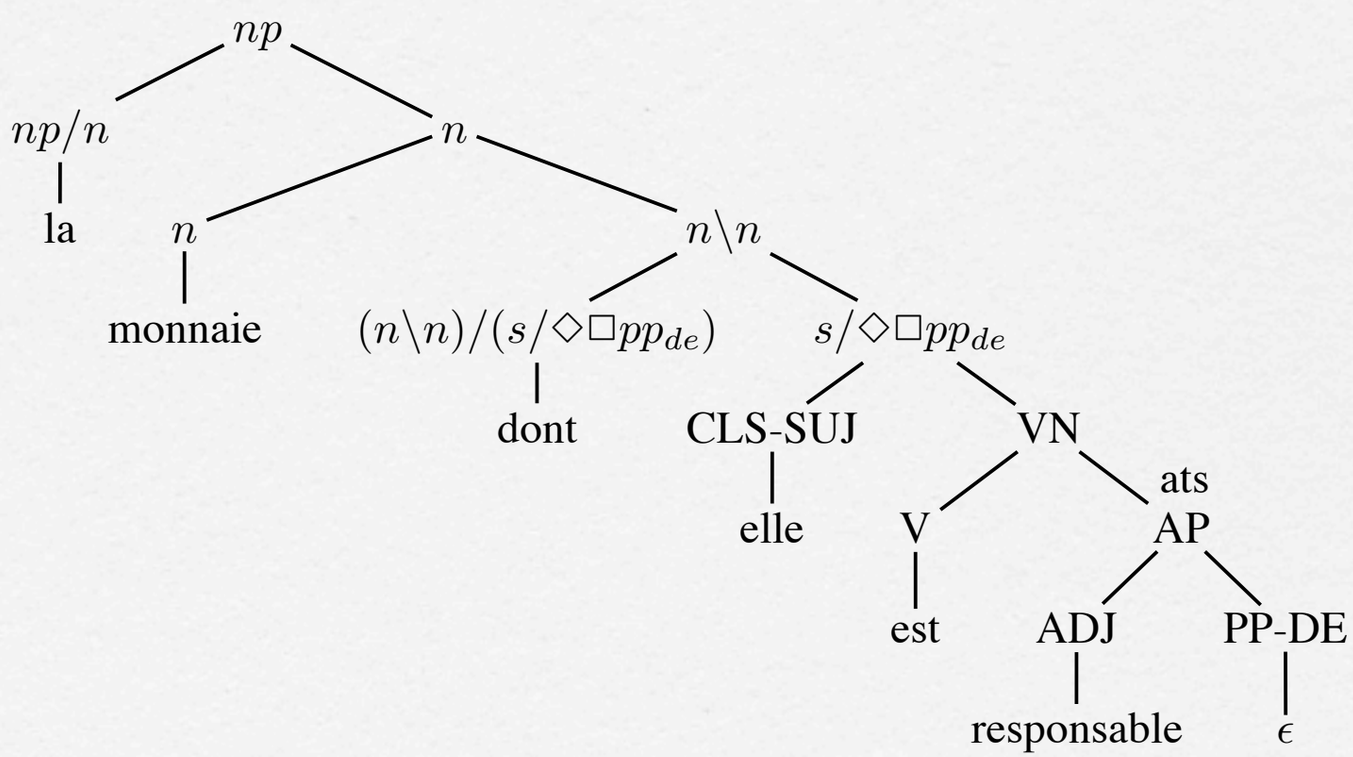
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



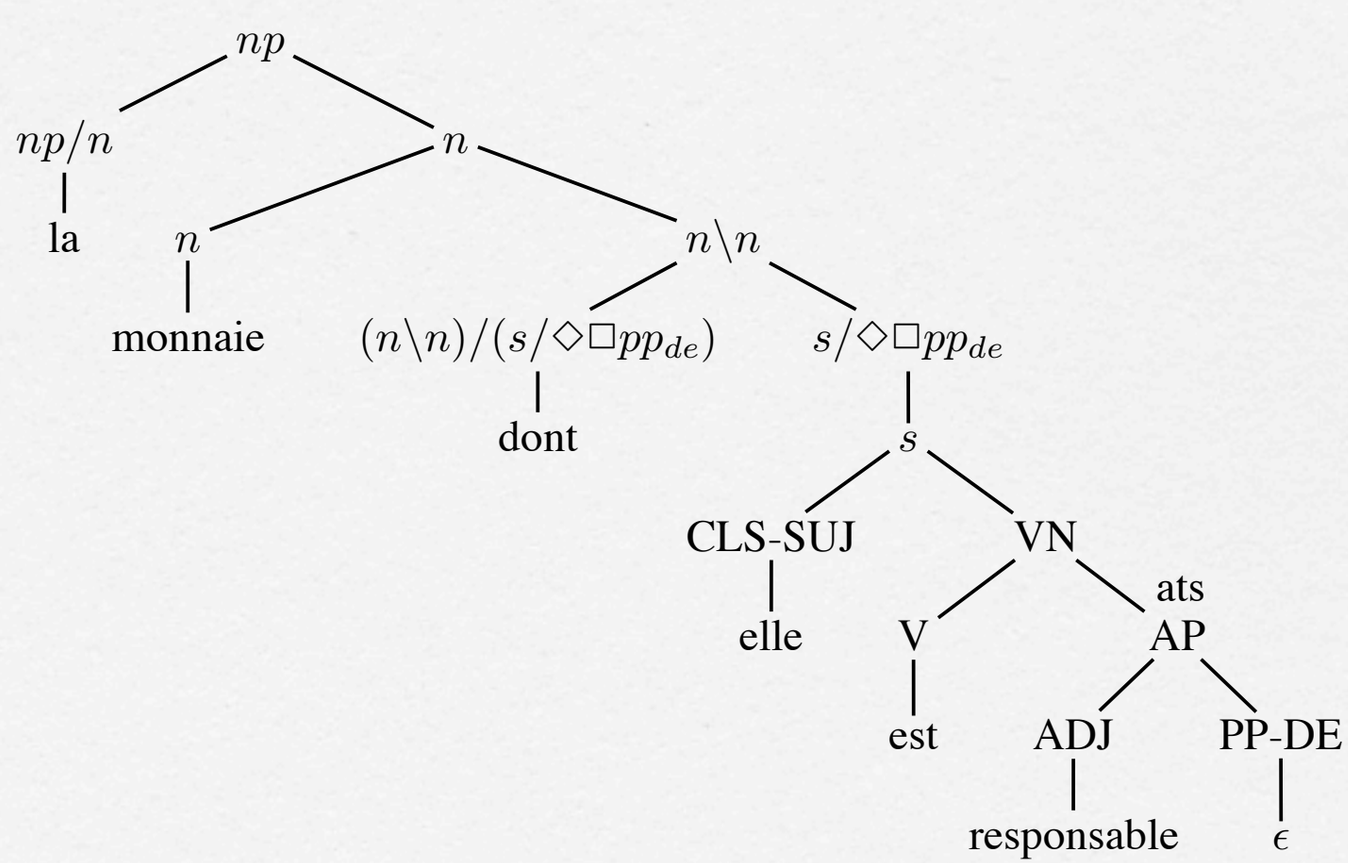
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



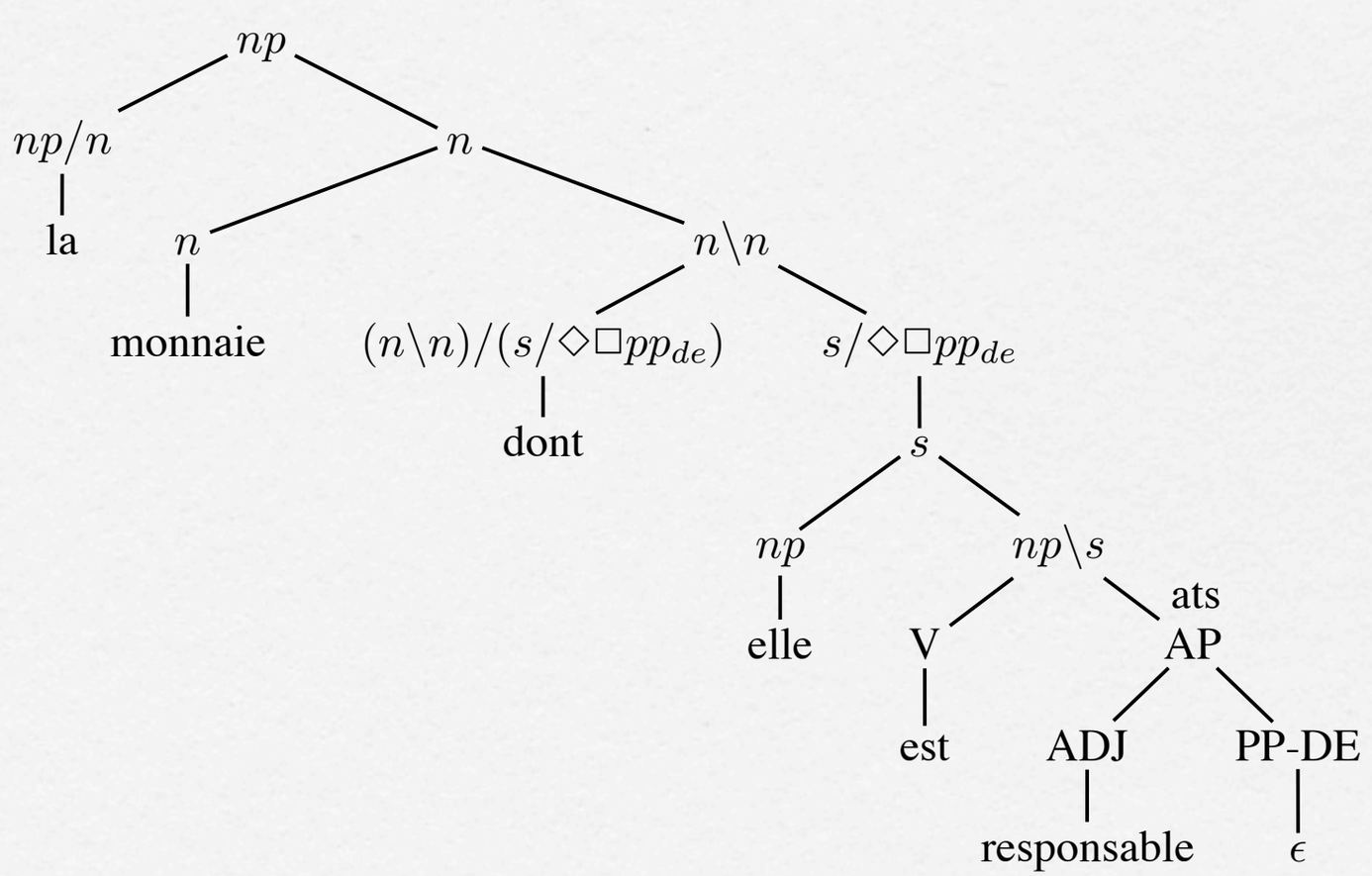
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



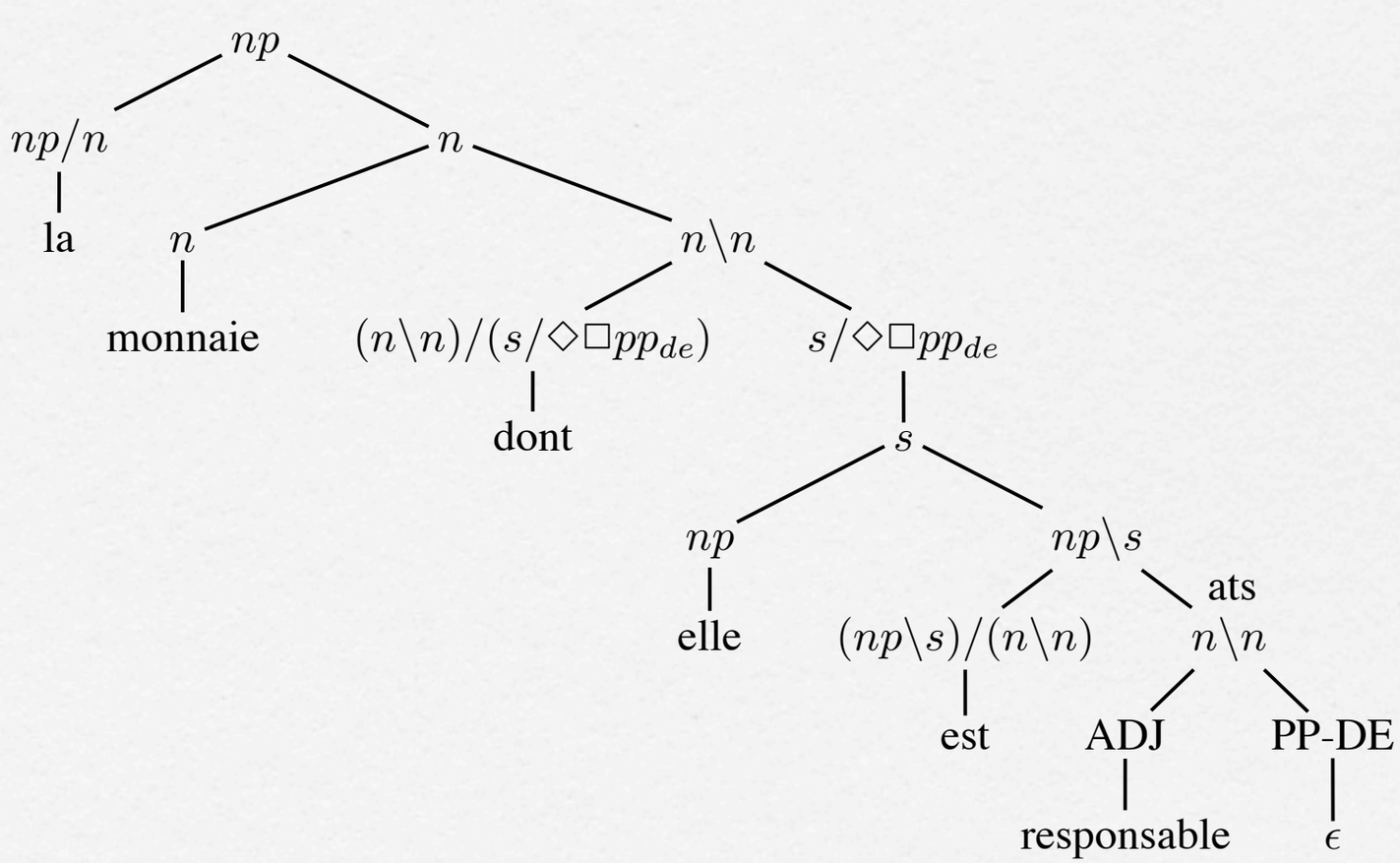
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



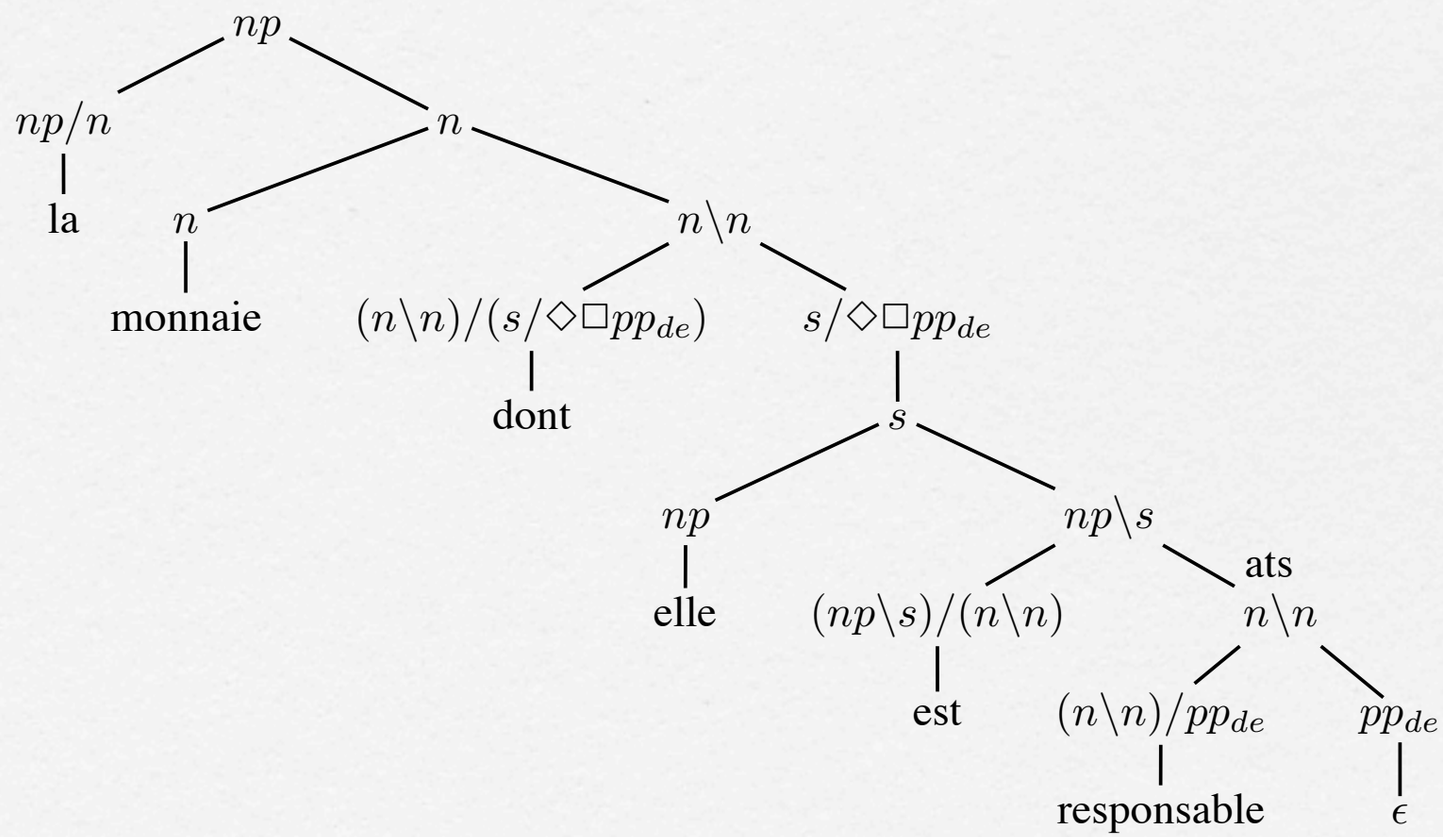
The extraction algorithm

1. Binarize the annotation
2. Assign formulas



The extraction algorithm

1. Binarize the annotation
2. Assign formulas



The extracted lexicon

- The extraction algorithm has been run on the 12.440 sentences (371.033 words) of the Paris VII annotated treebank.

- A total of 815 different formulas have been assigned to the words in the corpus.

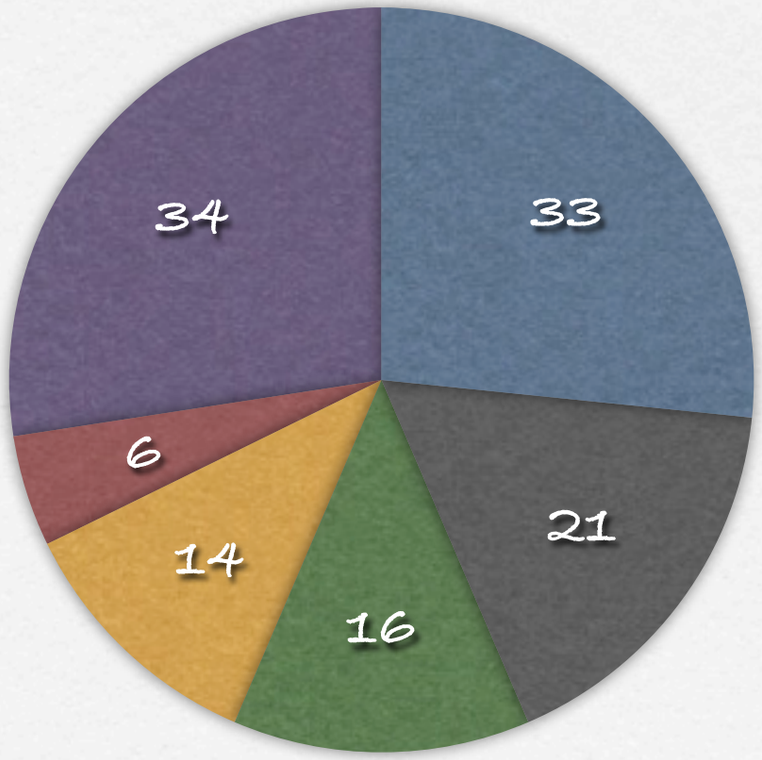
- Many frequent words have a large number of assigned formulas, eg. “et” (71) “comme” (42), “est” (37), “qui” (14).

The extracted lexicon

- Formula assignments to the present tense verb form “fait”

- 124 occurrences in the corpus with 19 different formulas assigned to it.

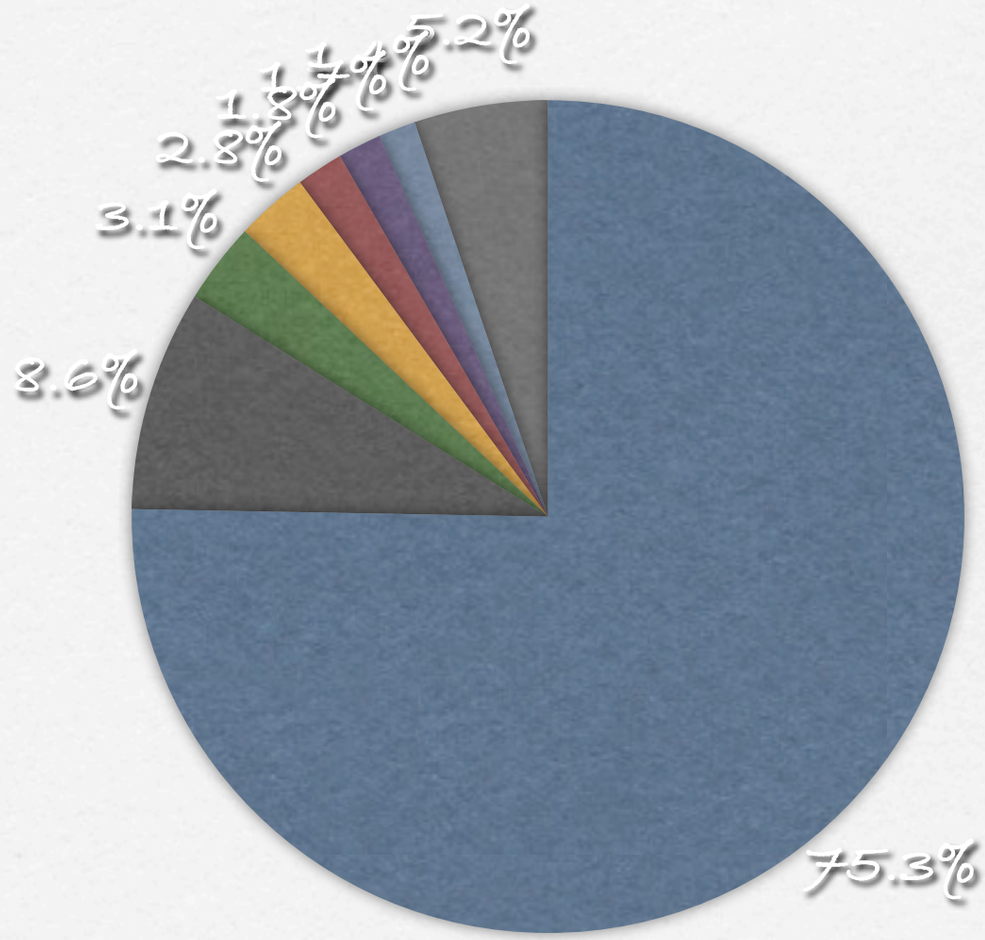
- (np\s)/np
- ((np\s)/pp_de)/np)
- (np\s)/(np/s_inf)
- ((np\s)/pp_a)/np
- ((np\s)/np)/(np\s_inf)
- other



The extracted lexicon

- Formula assignments to the comma ","

- 21,398 occurrences, 58 different formula assignments



- no formula
- (np\np)/np
- (n\n)/n
- (np\np)/n
- (s\s)/s
- ((np\s)\(np\s))/(np\s))
- ((n\n)\(n\n))\(n\n))
- other

The extracted lexicon

- The large number of assigned lexical categories becomes a hindrance to parsing.

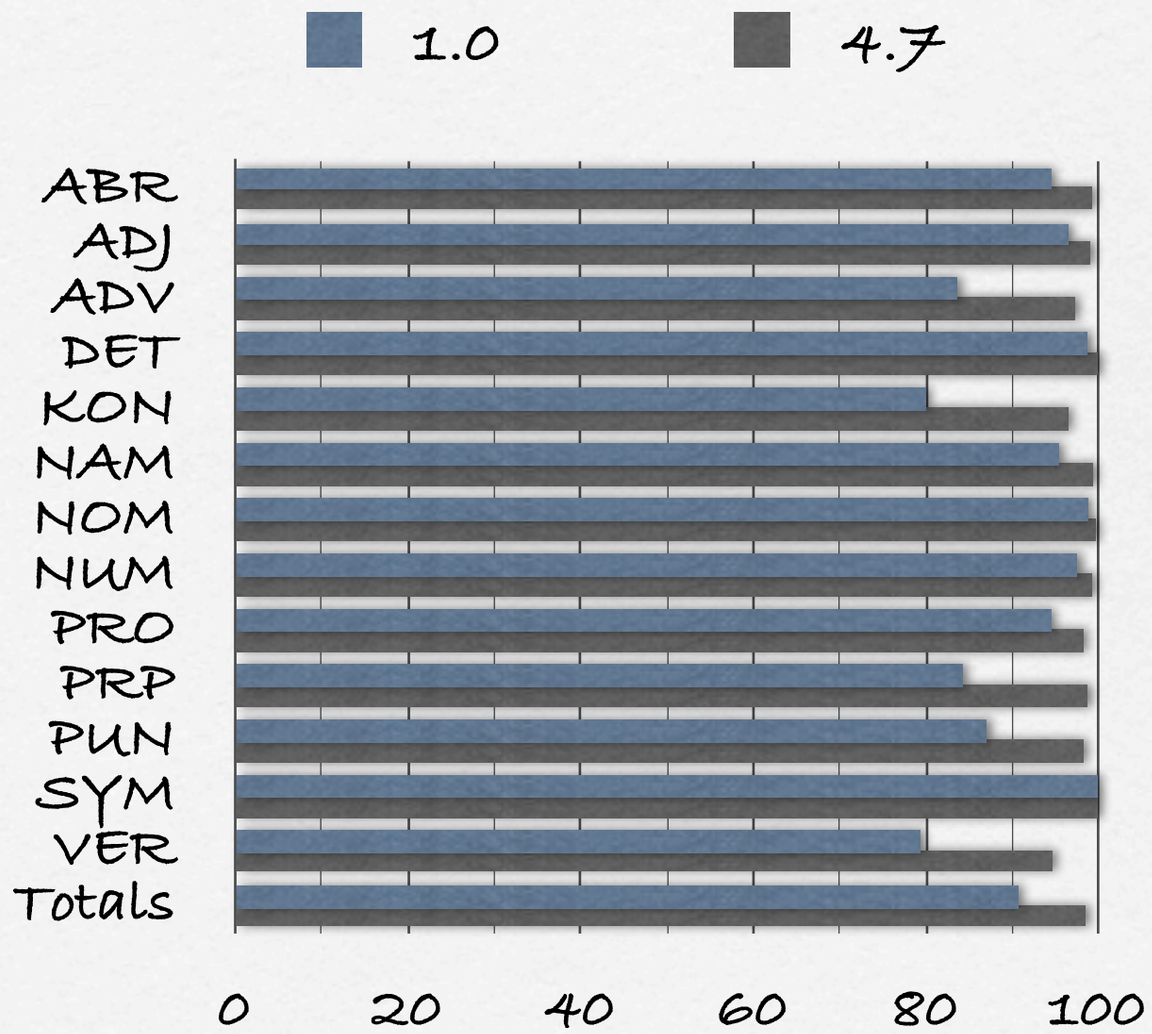
- A well-known solution to this problem is to use a supertagger which assigns each word-POS pair its most likely lexical category based on parameters estimated from the local context.

- By assigning each word all supertags within a certain probability window of the most likely supertag, we can increase the coverage.

Evaluation

- Supertagger performance is 90.7% for a single tagger and 98.4% for a multitagger which assigns (on average) 4.7 supertags to each word.

- this is the same level of performance as the best supertaggers for English.



Conclusions

- A wide-coverage type-logical grammar for French has been developed with semantic applications in mind.
- Evaluation using a supertagger gives state-of-the art performance
- Model files are distributed under the LGPL license.

Thank you!

Richard.Moot@labri.fr
http://www.labri.fr/perso/moot