

# RoDia: Fostering Language Diversity in One Corpus

Victoria Bobicev Cătălina Mărănduc

## Abstract

We present a corpus with rich morphological, syntactic and partially semantic annotation. Its main characteristics are the large variety of non-standard texts and several types of annotation.

The creation of this corpus pursues several objectives:

- (1) a better coverage of linguistic diversity of Romanian language;
- (2) diachronic analysis of Romanian;
- (3) creation of a gold standard annotation for various types of Romanian texts which permits;
- (4) creation of robust machine learning models for various types of annotation.

Nr.	Format	Sentences	Tokens
1	UAIC syntactic XML	32,753	671,235
2	UD syntactic CoNLLU	26,225	572,436
3	UAIC semantic XML	5,566	99,341

Table 1: Volume of the corpus annotated in each of the three formats: UAIC syntactic, UD syntactic and UAIC semantic.

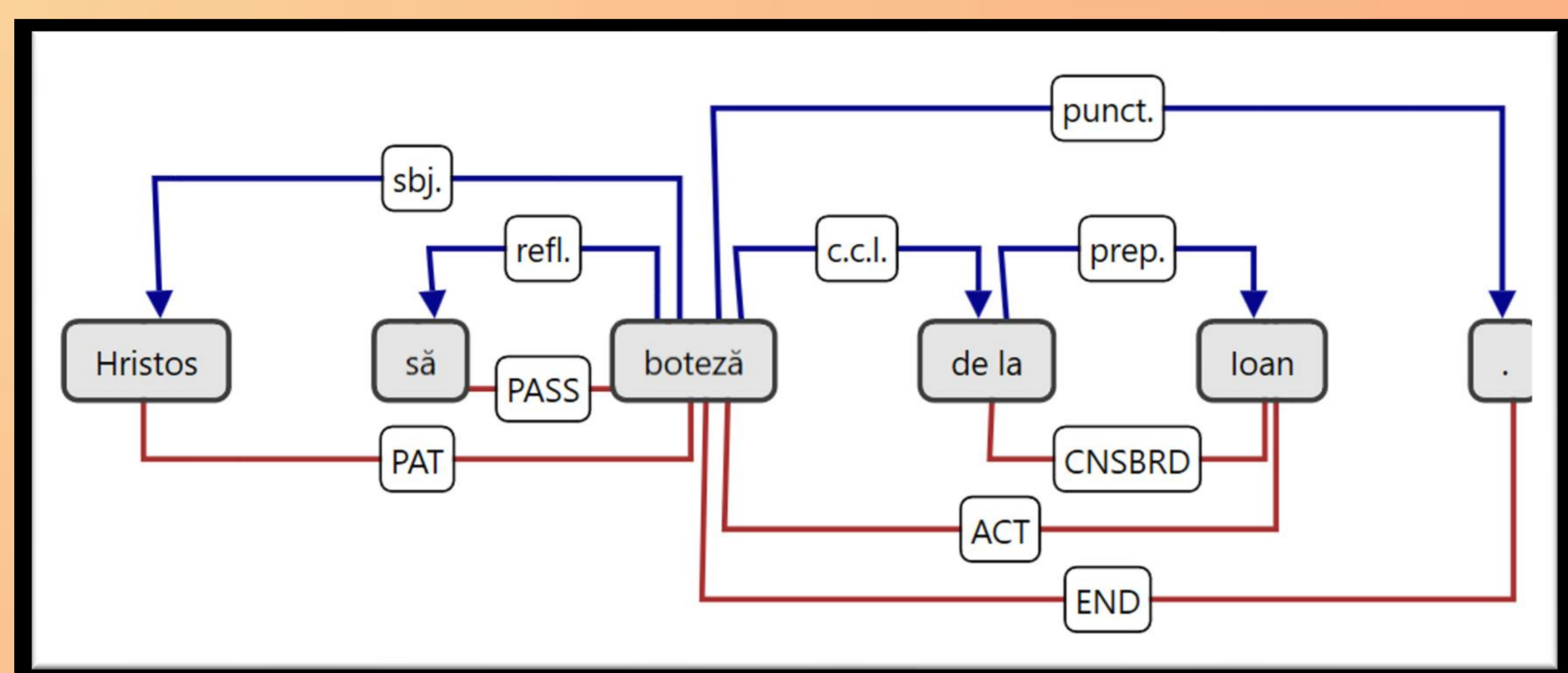


Figure 1: Comparison of the syntactic and semantic annotation of the sentence "Christ was baptized by John" (approximate translation).

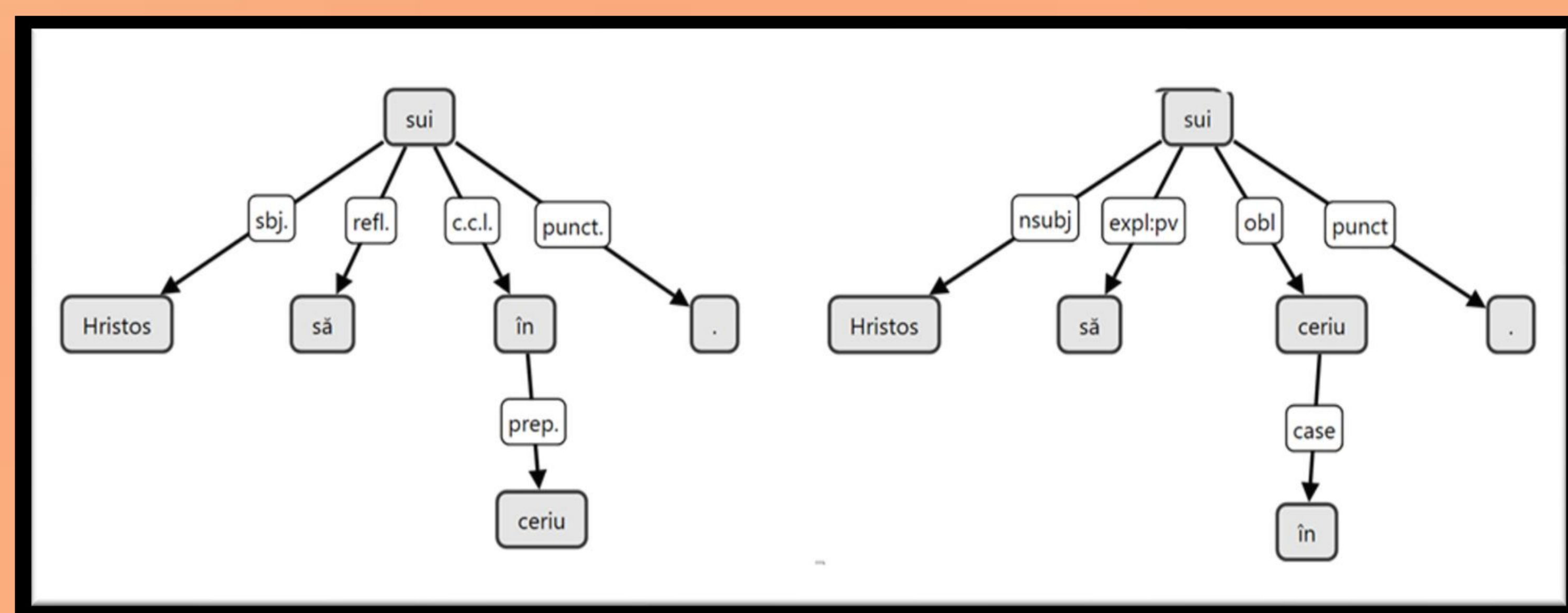


Figure 2: Christ ascended into heaven (ACTS\_1.9.content), in UAIC and UD conventions.

An example of UAIC syntactic annotation in xml format:

```
<sentence id="15" parser="Malt parser" user="ugla" date="2020-34-13">
  <word id="1" form="Unde" lemma="unde" postag="Rw"
  head="4" chunk="" deprel="c.c.l." />
  <word id="2" form="să" lemma="să" postag="Qs" head="4"
  chunk="" deprel="part." />
  <word id="3" form="să" lemma="sine" postag="Px3--a-----w"
  head="4" chunk="" deprel="refl." />
  .....
```

An example of TREEOPS rule for the transformation syntactic -> semantic annotation

```
//word[@deprel='coord.' and (@lemma='și' or
@lemma='nici')]/@deprel => changeAttrValue('CNCONJ')
```

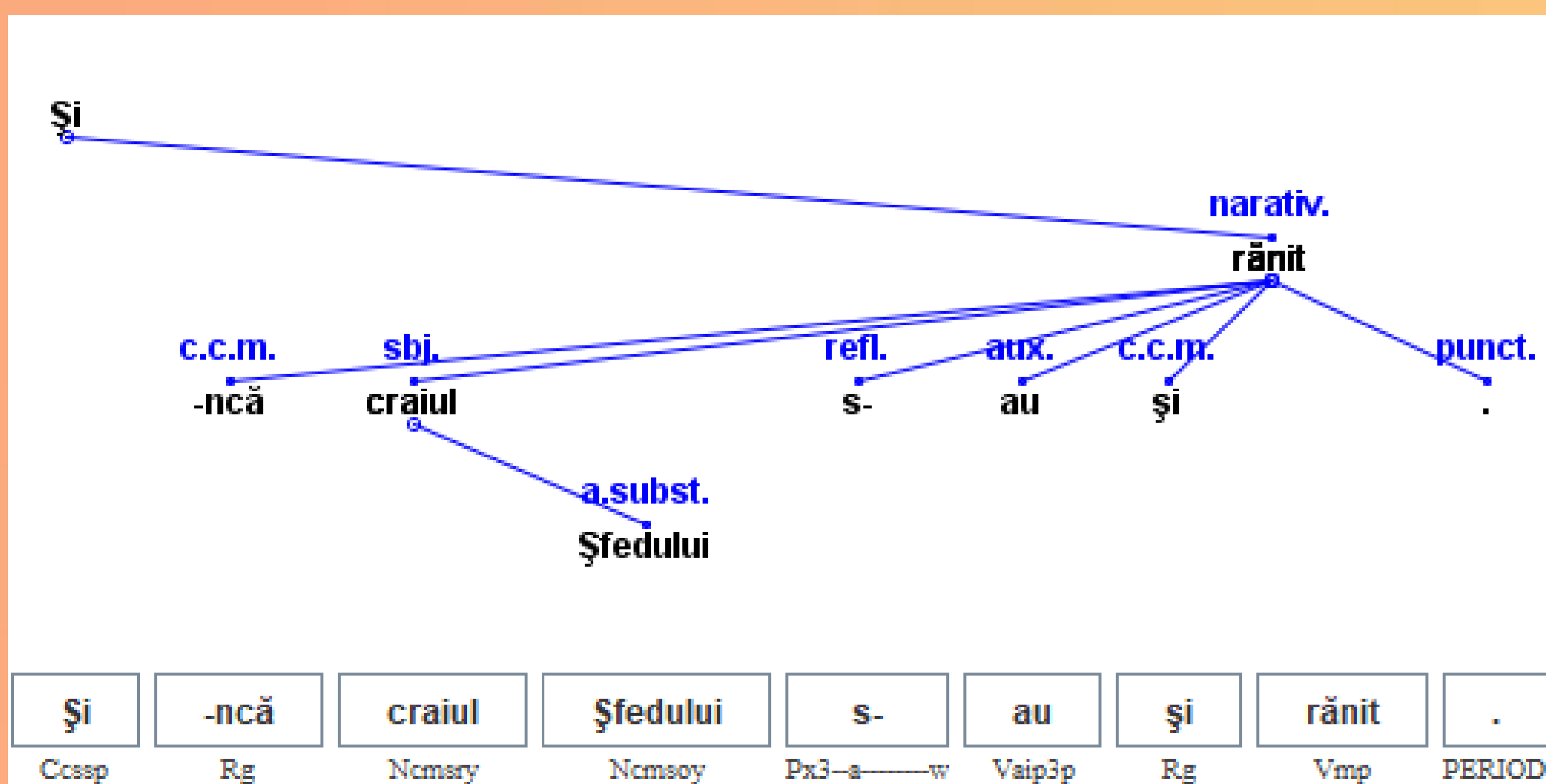


Figure 2: An example of the annotation interface TreeAnnotator (And also king of Sweden has been wounded).

An example of UD syntactic annotation in conllu format:

```
# sent_id = test-4
1 Avraam Avraam PROPN Npmsrn
Case=Acc,Nom|Definite=Ind|Gender=Masc|Number=Sing 2
nsubj _ ref=MATT 1.2
2 născu naște VERB Vmis3s
Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 0
root _ ref=MATT 1.2
3 Pre pe ADP Spsa AdpType=Prep|Case=Acc 4 case _
ref=MATT 1.2
4 Isaac Isaac PROPN Npmsrn
Case=Acc,Nom|Definite=Ind|Gender=Masc|Number=Sing 2
nmod:pmod _ ref=MATT 1.2
.....
```

An example of TREEOPS rule for the transformation UAIC syntactic -> UD syntactic annotation

```
if word1[@id="x", @postag="Sp*"] and word2[@head="x"]
then
word1/@head $\gets$ word2/@id
word2/@head $\gets$ word1/@head
foreach remaining wordN[@head="x"]
wordN/@head $\gets$ word2/@id
```

## Conclusion

The main aim of our work is the creation of the gold standard corpus to be used for future training of part of speech taggers and syntactic parsers; its volume should be enough for reliable parsing with minimum errors.

On the other hand, we need good annotation tools for faster corpus creation. Thus, our goals are interdependent: the corpus creation is dependent on the tools and the tools need a corpus for their training.

Given the rapid progress in language technology we believe that we can find and adapt a pipeline of tools that could help us expand our corpus faster and include a wider variety of documents in the corpus.