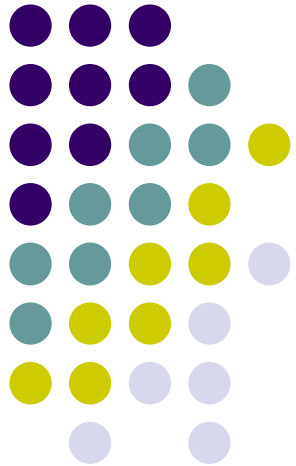
Named Entity Recognition with NooJ



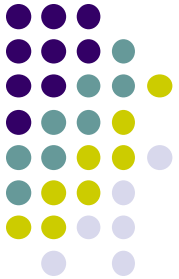
Aleksandar Petrovski

International Slavic University, Sveti Nikole

North Macedonia

Zagreb, NooJ 2020

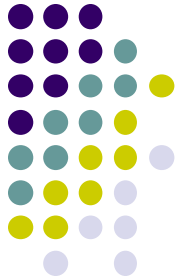
What are Named Entities?



Named entities, in a narrower sense, are concrete or abstract elements of the real world, including people, organizations, companies, places, etc., e.g. *Bill Gates*, *Google*, *Zagreb*.

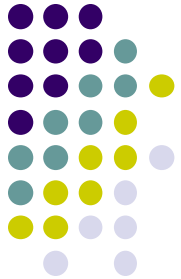
In a broader sense, the expressions related to time, space, quantity, such as September 11, 80 EUR, are considered as named entities too.

Why is Named Entity Recognition (NER) Important?



- Understanding documents
- Information retrieval - IR
- Machine Translation – MT
- Question Answering - QA
- Speech Recognition

NER Techniques

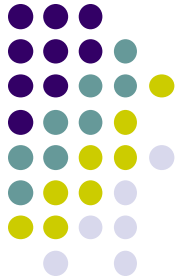


- Grammars (higher precision, but at the cost of lower recall, a team of computer engineers and computer linguists required, expensive)
- Statistical models (requires a large amount of manually annotated machine learning data, expensive, often only for a specific domain)

Suggested Solution

- NooJ linguistic environment
- NooJ morphological lexicons
- NooJ grammars (morphological and syntactic)

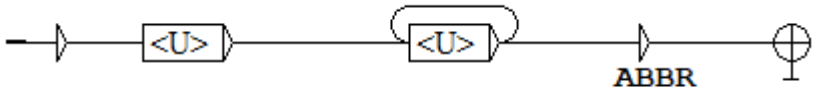
Macedonian Morphological Lexicons



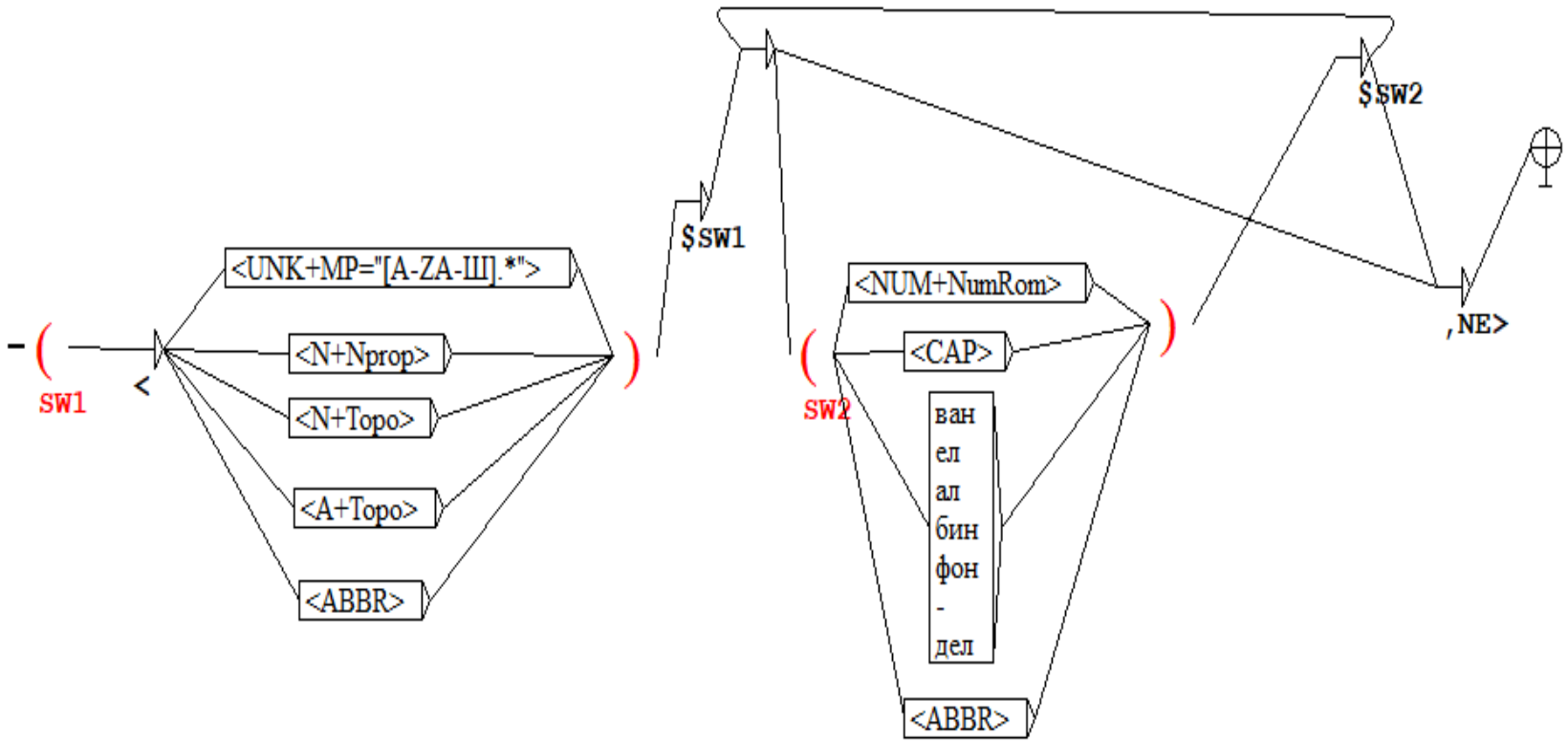
- Basic morphological lexicon (89,026 lemmas, 1,480.201 word forms)
- Lexicon of toponyms (1,398 lemmas, 40,246 word forms)
- Lexicon of names of persons and companies (5,422 lemmas, 157,321 word forms)
- Lexicon of derived adjectives from verbs, with suffixes -chki and -bilon (12,073 lemmas, 281,488 word forms)
- Lexicon of participles derived from verbs (19,552 lemmas, 1,252,328 word forms)
- Lexicon of compound units consisting of multiple words (784 lemmas, 6,289 word forms)

Morphological Grammars

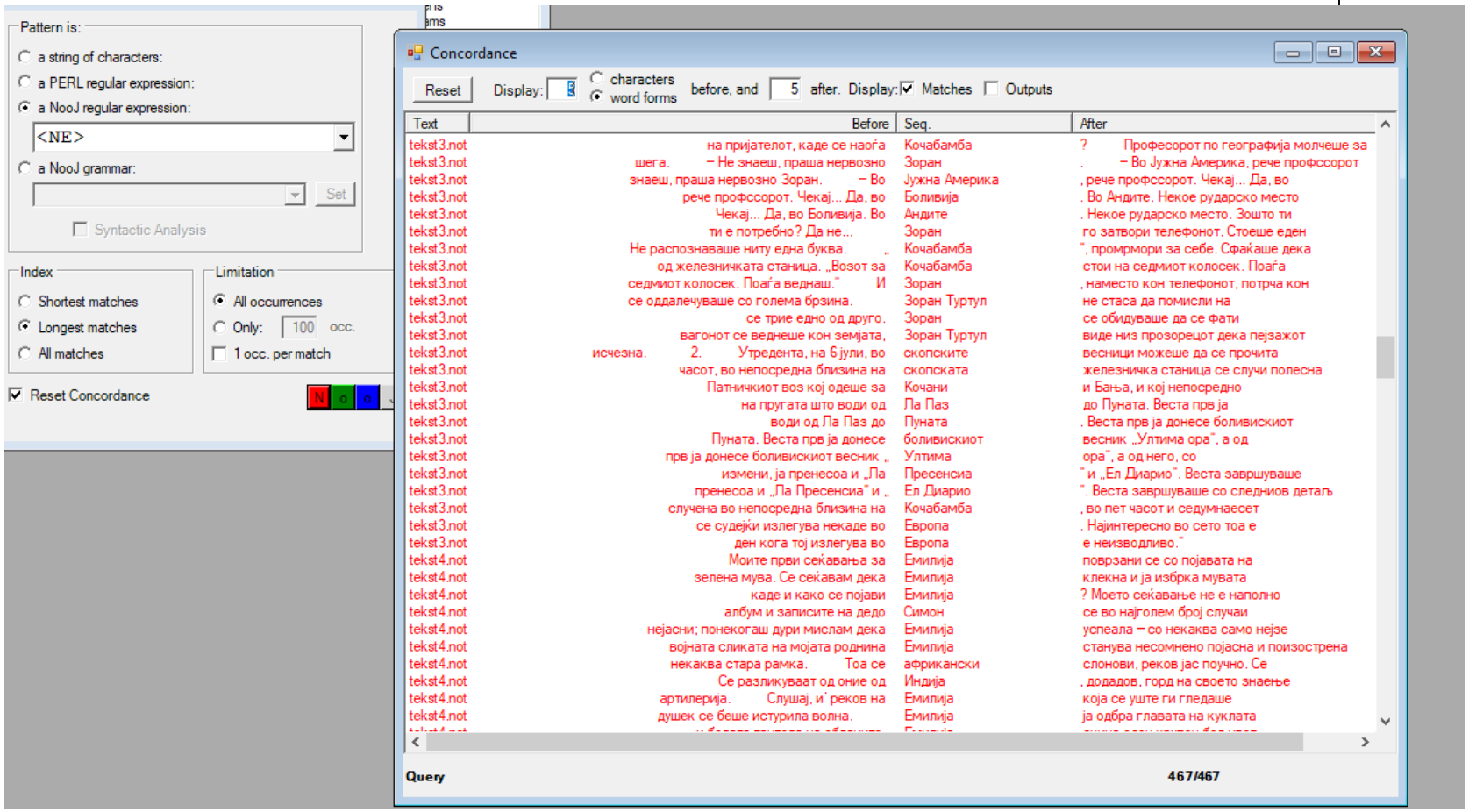
- Negative adjectives and adverbs
- Prefixed verbs
- Diminutives
- Augmentatives
- Roman numbers
- Abbreviations



Syntactic Grammar



Results


Pattern is:
a string of characters:
a PERL regular expression:
a NooJ regular expression:
<NE>
a NooJ grammar:
Set
Syntactic Analysis
Index
Shortest matches
Longest matches
All matches
Limitation
All occurrences
Only: 100 occ.
1 occ. per match
Reset Concordance
Concordance
Reset
Display: 3
characters
word forms
before, and 5 after. Display: Matches Outputs
Text
Before
Seq.
After
Query
467/467

Conclusion

In order to recognize the named entities, the program environment NooJ can be used, with the help of several language resources for Macedonian: morphological lexicons, morphological and syntactic grammars. A syntactic grammar that recognizes named entities in texts exploiting the capital letters rule is the basis on which their recognition is based. All of these language resources are applied to a small corpus, and the system detects 2.18% of named entities from the total number of word forms.

The system does not detect multi-word named entities, in which only the first word starts with a capital letter. This problem can be solved by developing new syntactic grammars.