Enriching Universal Dependencies Treebanks with Romanian Samples from Social Media

Diana Höfels

HS Tools and Resources for Low-resource Languages
SoSe 2023

Romanian Language

- Is it a low-resource language?
  - No
- Is there a genre-wise gap in the availability of UD treebanks for Romanian?
  - YES :)

Romanian Language

- Romanian is a Romance Language:
  - Evolved from Vulgar Latin spoken by Roman colonists in ancient Dacia.
  - Part of the same language family as French, Spanish, Italian, and Portuguese.
- Unique Alphabet and Diacritics:
  - Utilizes the Latin alphabet with distinct Romanian characters.
  - Diacritics include ”ă,” ”â,” ”î,” ”ș,” and ”ț” to represent specific phonemes.

Progress in UD Treebanks Annotation

1. Data Source: Coroseof Corpus (Hoefels et al. 2022)
   - Samples collected from Twitter
   - Sexist and offensive language
2. Progress so far ...
   - Annotated 50 Sentences (POS Tags and Dependency Relations)

Challenges

1. Rich morphology, word forms having multiple possible analyses, challenging to choose the correct POS tags and dependency relations for some words.
2. Social media language: non-grammatical, informal, slang, abbreviations, emojis, code-switching, ellipsis, no punctuation,
3. Word order flexibility: flexible word order compared to other languages, more difficult to determine the syntactic relationships between words in some cases.

Challenges

- Multi-word expressions:
  - *A da nas în nas*
  - Literal Translation: To give nose to nose.
  - Meaning: To bump into somebody.
- Word order flexibility: flexible word order compared to other languages, more difficult to determine the syntactic relationships between words in some cases.
  - Băiatul citește cartea.
  - Citește cartea băiatul.
  - Translation: The boy reads the book.

References

Hoefels, D. C., Çöltekin, Ç., and Mădroane, I. D. (2022). CoRoSeOf - an annotated corpus of Romanian sexist and offensive tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2269–2281, Marseille, France. European Language Resources Association.