Querying CoRoLa with KorAP: feedback from users

Verginica Mititelu

ICIA
Some previous remarks

• CoRoLa was a welcome initiative among the linguists studying Romanian
• For some of them – it answers older needs
• For others – still something new, uncontrollable, treated with reluctance or completely disregarded
Who are the users of CoRoLa

- Linguists
- Language engineers
- From Romania
- From abroad
How many are they?

• ?
• A counter would be useful for this
Have we advertised it enough?

• Definitely NOT!
How did we get feedback?

• Mainly by asking for it
• From the discussion sessions following the presentations we made
• Very few emails got through the contact form available on the website
Positive aspects (KorAP)

- Fast
- Robust
- The possibility to query it by combining different levels of annotation (token, lemma, morphology)
- The use of regular expressions
To improve: CoRoLa website

• Need for the list of occurring words with a link to their occurrences
• The server is sometimes down
To improve: corpus

- Texts quality <=:
  - Diacritic restauration
  - Automatic conversion from pdf to txt
  - Texts without punctuation: e.g. the file *colectie de fraze din wikipedia in limba romana*
  - Existence of **two orthographic norms** (â/î, sunt/sînt)

- Corpus dimensions: not large enough to reflect some tendencies in the language use!
  - Discussion: “problemelor acestea” N..poy Dd3.po.*

- Tokenization
  - Users are not aware of the conventions: e.g. *pentru că* is one single token, although two words
  - The same string of words is tokenized either as one token or as more tokens: *pentru că, ceea ce, în urmă, de la, în timp ce* etc.
  - How can we search for the unique token variant?

- POS-tagging
  - Need to be translated in Romanian
  - Can be too inclusive: e.g.: Rw
  - Contrary to the linguistic facts: din=Spsa, prin=Spsa, primprejurul=Spsg etc.
To improve: KorAP

• Creating the virtual corpus:
  – Unclear: which attributes from the list are applicable to CoRoLa
  – Unclear: what values these attributes can have in the metadata

• Writing the query:
  – Presupposes becoming familiar with the annotation principles
  – Drop-down lists with values of attributes used in a query would be useful
  – Preference for a certain encoding of quotes: NO: “”, YES: ""
  – Inconsistent/counter-intuitive use of quotes:
    [drukola/orth=când | drukola/orth=cînd] o [drukola/m="msd:v.np"]
  – Impossibility of exploiting existing punctuation
    • Returning results disregarding the sentence limits
    • Unsolved ambiguity: e.g. Rw
To improve: KorAP-search

• Negation in regexps: works intermittently:
  Q:
  [drukola/m="msd:nc..o.*"] [drukola/m!="msd:s.*"] [drukola/m="msd:nc..rn.*"]
  R: (...) colegelor de cameră

  Q: [drukola/m="msd:ncfso.*"]
  [drukola/m="msd:d..fsr.*"] [drukola/m!="msd:n.*"]
  R: (...) reglementări, această indemnizăție

• Impossible to save the results
To improve: KorAP display

- Order of results: blogs always come first
To improve: KorAP display

- Metadata in different order from one snippet to the other:
To improve: KorAP display

• Splitting words preceded by certain punctuation signs: , : “ (
To improve: KorAP display

• Useful: hovering over a POS tag displays the unabbreviated form of the part of speech
• The morphological information is presented in an unexpected order for a linguist: alphabetic
To improve: KorAP interface

• Need to be localized
• Need to have a more comprehensive manual
Conclusions

• CoRoLa is a valuable resource for Romanian.
• KorAP offers (potential) users the possibility to explore it and benefit from it.
• There is still place for improvement and this will require further collaboration between the DruKoLA partners.
• What is your experience of working with CoRoLa under KorAP?