DEVELOPING A POLYSYNTHETIC LANGUAGE CORPUS: PROBLEMS AND SOLUTIONS

Timofey Arkhangelskiy & Yury Lander

West Circassian (Adyghe)

- Spoken by about 500.000 people in the North Caucasus, Turkey and other Near East countries
- Written (novels, poems, newspapers and journals)
- Highly polysynthetic
  - A word may convey much information that is usually expressed syntactically in Standard Average European languages.
  - A word may have a highly complex structure.

qə-gʷə-rə-ʔʷe-**ŝʷa-ʁ**

DIR-heart-LOC-say-SEEM-PST

'It seemed that s/he understood that.'

  - Variable affix order, morphological recursion etc.

The Circassian corpus project

- Supported by Russian Foundation for Basic Research
- Involves Timofey Arkhangelskiy, Irina Bagirokova, Yury Lander, and Georgi Moroz
- Is based on the modified UniParser platform

The Circassian corpus project

- Is intended to include as much morphological information as possible
- Allows search based on words, morphemes and their combinations
- The first open polysynthetic corpus of this kind

Specific issues related to polysynthesis

- Tokenization
  - Problems related to the difficulties in demarcating between syntax and morphology
- Part-of-speech tagging
  - Wide distribution of affixes
  - Nominalizers and verbalizers within the same word
- Lemmatization
  - Very productive affixation
  - on a par with frequent non-compositional combinations of affixes
- The importance of morphological information
  - Searchable glossing

Web interface and query language

- The specific query language based on morphological glossing
- Query on glossing may be combined with a query on the bag of tags