From Kyrgyz Internet Texts to an XML Full-form Annotated Lexicon: a Simple Semi-automatic Pipeline

Loïc Boizou (1)
Dinara Mambetkazieva (2)

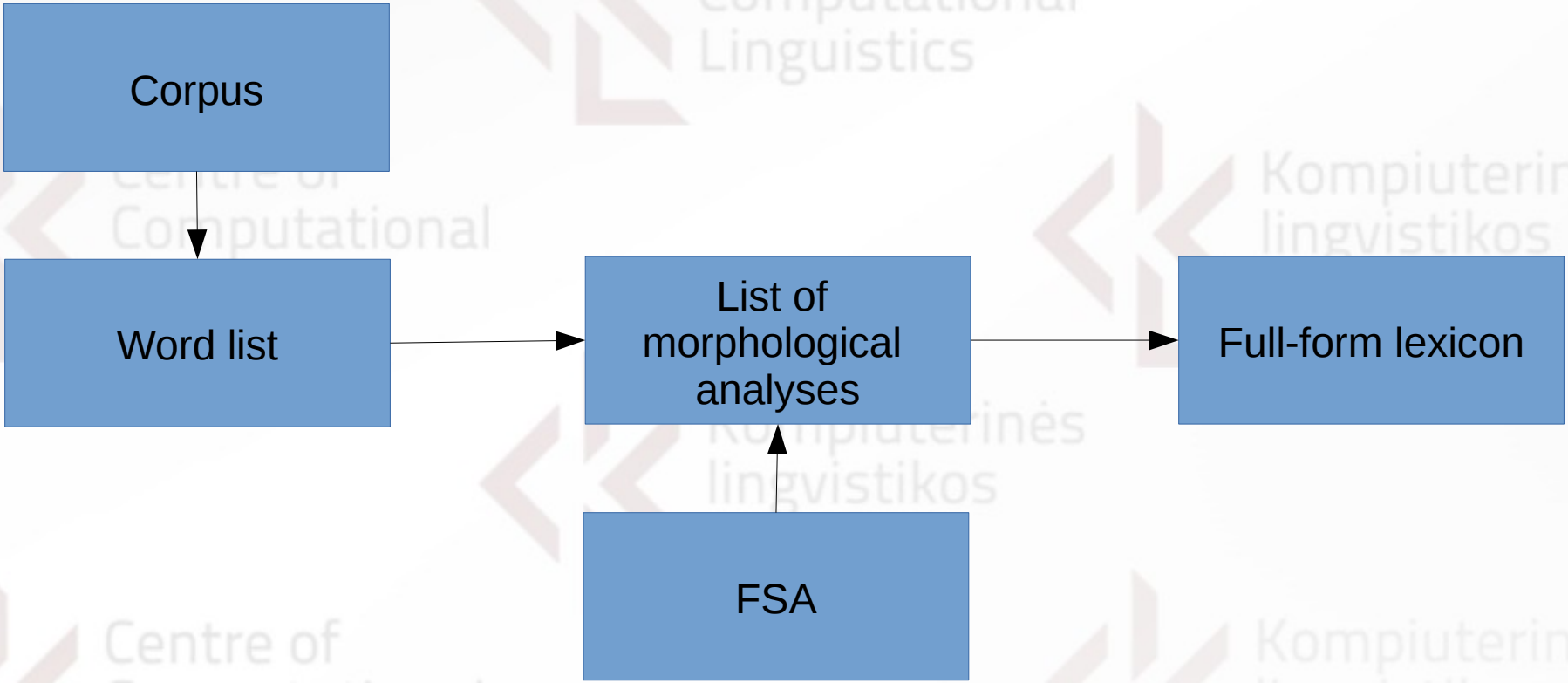
Vytautas Magnus University
(1) Centre of Computational Linguistics
(2) Institute of Foreign Languages

Introduction

- Despite its official position in Kyrgyzstan, Kyrgyz language still lacks resources, especially free resources that could be shared to develop applications.
- Nothing really new in the data processing presented in this work.
- Attempt to generate a free Kyrgyz resource as a way to foster future cooperation.

The pipeline

- Remark: Automatic components in Haskell



I) Word list extraction

The Corpus (1)

- Full texts (except *Manas*) available on line.
- Standard written language.
- 4 groups:
  - Literary texts;
  - News texts;
  - Institutional texts (universities, companies, state institutions);
  - Wikipedia articles.

The Corpus (2)

- Size 170 texts, 1.6 million running words.
- Issues:
  - Small corpora.
  - Improper balancing (1/2 literature, overweight of some texts/domains).
- **But**: diversity seems sufficient to capture the basic lexicon.

The word list

- Filtering & sorting.
  - Strings of Kyrgyz Cyrillic letters (+ digits).
  - Russian words left…
  - Latin-Cyrillic mixed strings removed.
- Close to 130,000 distinct word forms.

II) Morphological analysis

Structure

- 3 steps:
  - Pre-processing.
  - Analysis with a finite-state machine.
  - Post-processing.

Pre-processing

- Double-sound letters *ю, я, ё* (and *е* after vowel)are replaced by equivalent letter sequences *йу, йа, йо (and йе)*.
- Easier for morphemic segmentation:
  - ex. *коюп* → *койуп* (morphologically *кой-уп*)

The finite-state machine

- Raw FSA.
- Suffixes only (intensive adjective not treated).
- Analysis from the end of the word until failure. Longer stem are preserved as alternative interpretations (parse tree).
- Guesser-style → all plausible segmentations are provided.
- No disambiguation at this stage.

The FSA (1)

- Stored as simple Unicode text file.
- Each transition: starting state, next state, input string, grammar features.
- In general, 1 transition = 1 morphemic form.
- About 1000 transitions.

The FSA (2)

- Sequence *possessive* + *case* (often irregular) are treated together a single transition.
  - *эл-име* / *эл-ине* (vs. *эл-ге*, *эл-и*)
- According to the result of the analysis, the stem is marked as verbal, nominal or both nominal and verbal.
  - Adjectives are distinguished from nominals only for few suffixes.

The FSA (3)

- Opposition between derivational and inflexional suffixes marked as a feature.
- But some suffixes are on the border…
  - Privative suffix -*сыз*, e.g. *карындаштарым**сыз*** “without my younger sisters” (utterance level), vs. *жумуш**суз**дук* “unemployment” (embedded in a lexical derivative)
- Such suffixes appear twice in the automaton.

Post-processing

- Generation of lemmas (removal of flexional suffixes), selection of relevant features.
- Letters sequences *йу, йа*, *йо, йе* are reversed back to *ю*, *я*, *ё, е* inside morphemic units.
  - *Койон + дон → Коён + дон*

III) Preparation of the lexicon

Filtering

- Too short stems are removed:
  - One-letter stems and two-letter stems ended in vowel (except *де* “to say” and *же* “to eat”).
- The simple FSA analysis left many cases of under- and over-stemming (Moral et al., 2014)
- Automatic removal of some under-stemming analysis (cautious approach...).

Manual correction

- The list of alternative morphological analyses is reviewed by a native speaker.
- Simple process,
  - ‘+’ sign before correct morphological interpretations;
  - No correct interpretation → Direct correction.
- Automatic suppression of blacklisted stems (marked with a ‘-’ sign).
- Further step: a bit of automatic disambiguation?

Generation of the XML lexicon

- Automatic conversion of entries.
- According to TEI P5 standard.
- Structure directly follows the model provided by Budin et al. (2012).

The UD features

- TEI P5 let you choose the way you define grammar features.
- Universal Dependency defines part of speech and features descriptions.
- Available works about UD use for Turkish (Çöltekin, 2015, Eryigit et alii, 2016).
- Most features have natural counterparts, but some issues remain with the verb forms and modo-temporal categories (Kaşıkara, 2015).

Chosen UD verbal features: finite forms

- жазды : Tense=Past, Aspect=Perf
- жазган : Tense=Past, Aspect=Imp
- жазучу : Tense=Past, Aspect=Iter
- жазыптыр : Tense=Past, Evident=Nfh
- жазат : Tense=Pres (although it often expresses future)
- жазар : Mood=Pot
- жазса : Mood=Cond
- жазсын : Mood=Imp (although Mood=Opt may be better)

Chosen UD verbal features: non-finite forms

- жазуу : VerbForm=Inf
- жазган : VerbForm=Part (homonym of the finite imperfect form 3rd person)
- жаза : VerbForm=Conv, Tense=Pres (Aspect=Imp might be an option)
- жазып : VerbForm=Conv, Tense=Past (Aspect=Perf might be an option)

Final remarks

- Lexicon size: about 20,000 lexemes?
- More shared resources.
- Standards → re-usability + comparability.
  - Common UD features (other presentations on this topic).
- A common internet space for Kyrgyz resources?
  - Possibly on a larger Turkic space.

References

- Baisa, V., & Suchomel, V. (2012). Large Corpora for Turkic Languages and Unsupervised Morphological Analysis, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12). Istanbul, Turkey.
- Çöltekin, C. (2015). A Grammar-Book Treebank of Turkish, Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14). Warsaw, Poland.
- Eryigit, G., Gokirmak, M., Nivre, J., Sulubacak, U., Tyers, F.M., & Çöltekin, Ç. (2016). Universal Dependencies for Turkish. COLING.
- Kaşıkara, H. (2015). Universal Dependency Representation of Turkish: The Challenge of the Verb, Master thesis. Uppsala University.
- Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval, Information Research, 19, 1.
- Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (pp 3-16). Springer/

Thank You for Your attention!

Questions, remarks, suggestions?