Building an Albanian Text Corpus for Linguistic Research

Besim Kabashi
Friedrich-Alexander-Universität Erlangen-Nürnberg
besim.kabashi@fau.de

In this talk we are going to present our on-going work to create an Albanian text corpus for linguistic research. The main properties of the work and the topics of our talk are:

[Size of corpus] The aim of this work is to collect approximately 100 million words (tokens) of texts written in Albanian and of transcriptions of spoken Albanian. The corpus size is comparable to that of the British National Corpus (BNC).

[Domains] At the moment, we have collected around 70 million tokens. The texts are mostly taken from the Albanian press (daily and weekly newspapers). Some texts, around 10 %, are from the belletristic domain. The remaining texts are books from different domains, like geography, history, economy or medicine.

[Newspaper texts] The texts taken from newspapers are collected from different domains and according to a large set of criteria, e. g. subtype of text (interviews, columns, etc.), gender (female or male author), text length, etc. The goal is to compile this part of the corpus as balanced as possible, i. e. to cover as many different aspects of text types and text producers as possible.

[Standard Albanian and dialects] The texts are standard Albanian texts. Some texts that are written in a dialect or in an idiolect (a language variety unique to an individual), or before standard Albanian was regulated in 1972, are collected as a separate part of the corpus and are marked correspondingly — this allows us to include or exclude these texts from our queries depending on the linguistic research task.

[Annotation of corpus] At the current stage, the corpus is not yet annotated, but we have developed a morphological analyzer (Kabashi, 2015) which covers Albanian inflection and the most frequent types of word-formation. A lot of geographical and personal names are also covered by the morphological component. As a next step, the corpus will be POS-tagged and morpho-syntactically annotated. A tagset for this is now available (Kabashi and Proisl, 2016).

[Working with the corpus] The corpus is not publicly available at the moment. We have also not decided on a user front-end to explore the corpus, yet, but we plan on testing a couple of popular choices, e. g. CQPweb or the NoSketch Engine. At the moment, we are using unix-tools and programming language scripts to query the corpus. For example, we can easily extract n-grams, concordances for some exact word-forms, segments or sequences of letters, words, or sentences.

[Further works] Constructing the corpus is a one-person work that goes very slowly forward. Nevertheless, we look forward to further extending the corpus-size to at least 100 million tokens. We will also try to come up with a way to make this corpus available to linguists who would like to work with it.

[Difficulties compiling the corpus] Getting the texts, getting different kinds of texts, their quality and their classification, are the biggest problems.

References

Kabashi, Besim (2015): *Automatische Verarbeitung der Morphologie des Albanischen*. XVIII, 211. Erlangen, FAU University Press, 2015. ISBN (Print): 978-3-944057-40-8. eISBN (Online): 978-3-944057-43-9. ISSN: 2198-8102.

Kabashi, Besim; Proisl, Thomas (2016): "A Proposal for a Part-of-Speech Tagset for the Albanian Language". [In:] *Proceedings of the tenth conference on International Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016.