Development of the First LRs for Macedonian: Current Projects

Ruska Ivanovska-Naskova

Faculty of Philology- University "St. Cyril and Methodius"
Bul. Krste Petkov Misirkov bb, 1000 Skopje, Macedonia
rivanovska@flf.ukim.edu.mk

Abstract

This paper presents in brief several ongoing projects whose aim is to develop the first LRs for Macedonian, in particular the raw corpus compiled by Prof. George Mitrevski at the Auburn University, the preparation for the compilation of a reference corpus for the Macedonian written language at the MASA (Macedonian Academy of Sciences and Arts), the first small annotated corpus of the Macedonian translation of the Orwell's "1984", the electronic dictionary of simple words created by Aleksandar Petrovski for the Macedonian module in the frame of the corpus processing system Intex/Nooj and the Morphological dictionary developed by the LTRC (Language Technology Research Center). Further we discuss the importance of the development of the basic LRs for Macedonian as a means of preservation and a prerequisite for the creation of the first commercial language products for this Slavic language.

1. Introduction

The Macedonian language belongs to the group of minority or, so-called, lesser-used languages that due to lack of funding, specialized human resources and a relatively small market for commercial language products is way behind the leading languages in the field of the computational linguistics. Still the creation of LRs for Macedonian is essential for its preservation because it will encourage linguistic investigations that include this Slavic language and will act at the same time as a starting point for the development of commercial language products.

We would like to present few ongoing projects that aim to develop the first corpora and electronic dictionaries for Macedonian.

2. Corpora

2.1 The first Macedonian online corpus

The first initiative was launched by Prof. George Mitrevski at the Auburn University, Alabama who has been working on the compilation of a Macedonian written corpus (Mitrevski). Its uncompleted version (one million words approximately) is consultable on the Web (http://omilia.uio.no/CE/mak/). It is a raw corpus developed with the IMS Corpus Workbench of the Institut für Maschinelle Sprachverbeitung at the University of Stuttgart. The corpus is made up of around 10 different types of texts mainly retrieved from the Internet. Each text is described with several parameters: its number in the database, title, author, genre, subject, publisher, date of publishing, date of its registration in the database, ISBN or other identification number, text format, information whether a sample or the entire text is included, number of words etc. The corpus can be used to build concordances for single words or groups of up to five words, collocations etc. The queries can be applied to the whole corpus or only to a group of texts selected according to the criteria included in the description of the texts. The future development of the corpus regards the size of the corpus (it is planned to reach ten million words) and POS annotation compatible with the Corpus Encoding Standard (CES) and the multilingual MULTEXT-East data set .

2.2 The MASA reference corpus of written Macedonian

The second initiative comes from the Research Center for Areal Linguistics at the Macedonian Academy of Sciences and Arts. Ac. Zzuzana Topoljinska and her team launched the idea related to the development of reference corpus for the Macedonian written language which is a part of a larger project at regional scale of the network of the Academies of South-Eastern Europe. The project is still in its preparatory phase when many features of the corpus and the strategy for its development are yet to be defined. Currently a team of linguists is working on the typology of authentic texts written in modern Macedonian that are going to be included in the corpus, their preparation and the creation of a referencing system. The planned degree of annotation is on morphological level (POS tagging). A team of engineers is working on the selection of tools that correspond to the nature of the Macedonian language and their adaptation for the organization and the treatment of the corpus. The actual compilation of the corpus is planned to start in the course of 2006. Similar initiative is to be undertaken by the Institute of the Macedonian Language "Krste Petkov Misirkov". (Venovska-Anteska, 2005).

2.3 The tagged Macedonian translation of Orwell's "1984"

This first small-size annotated corpus of the Macedonian language is part of the bilateral Macedonian-Slovene project "Gathering, annotation and analysis of Macedonian-Slovene language resources" in which participated several researchers (Zdravkova et al., 2005; Vojnovski et al., 2005). This is the first attempt to create Macedonian morpho-lexical resources conform to the guidelines of MULTEXT-East. The text of the corpus is the Macedonian translation of the Orwell's "1984". The first stage of the creation of the corpus was the scanning of the paper version of the text and its conversion into a digital format. The preprocessing stage also included segmentation, tokenization and compilation of a dictionary of the word forms which later were annotated. Each word form is associated with a morpho-syntactic description: part-of-speech tag (11 grammatical categories) and information about the corresponding

attributes (84 for the Macedonian language) and their values (134). The morpho-syntactic description is represented as a string of different characters. The word forms were semi-automatically classified: 60% automatically according to the inflection and the rest of the words manual. Each word is also associated with its lemma. This annotated corpus was used for learning of the TnT (Trigrams'n'Tags) which is an efficient language-non-specific statistical part-of-speech tagger suitable for training on large corpora. The tagger tested on the same corpus achieved an accuracy of 98.1%. The further work of this research group regards the finalization of the lexical lists through a rule-based lexicon, re-learning of the tagger and its testing on another text.

3. Electronic dictionaries

3.1 Intex/Nooj electronic dictionary of simple words

The second part of this paper will focus on the compilation of two electronic dictionaries. Aleksandar Petrovski is developing an electronic dictionary of simple words which is a starting point for the creation of the Macedonian module in the frame of the Intex, recently Nooj corpus processing system. (Petrovski, 2005; Silberztein, 2005) One of the main aims of Intex/Nooj is to allow the construction of formalized description of languages and apply them to large corpora compiled acoording to the needs of the user. The main linguistic resources of this development environment are the e-morpho-syntactical dictionaries and various types of grammars (inflectional, derivational, lexical, orthographical, syntactic, semantic etc.) represented as a set of graphs. Finite-state automata, finite-state transducers and other computational devices are used for the formalization of the linguistic phenomena. The system works through language-specific modules developed by several teams of researchers that can be upgraded and modified by the user. The level of elaboration of the module differs from language to language.

The core of each language module are the e-dictionaries that are conformed to the methodology promoted by the RELEX network: the first step is the creation of the dictionary of lemmas and corresponding flectional codes (DELAS) in order to automatically build the dictionary of all inflected forms (DELAF). Petrovski has started working on the Macedonian DELAS presented at the 8th Nooj workshop:

Figure 1. Table of the word groups, the categories and the codes (Petrovski, 2005)

The main source used for the creation of the lexical database was the Blaze Koneski's traditional dictionary of the Macedonian language. The dictionary was scanned and the errors were corrected. The basic tag set presented in Figure.1 is formed of 10 grammatical categories represented with corresponding codes. Each grammatical category is further described with certain number of attributes and values (ex. the nouns are represented with four attributes: gender with three values, number with four, case with three and definiteness with three values) At the moment of the presentation of the paper, the Macedonian DELAS consisted of 61.296 lemmas which produce 426.161 inflected forms as shown in Figure.2:

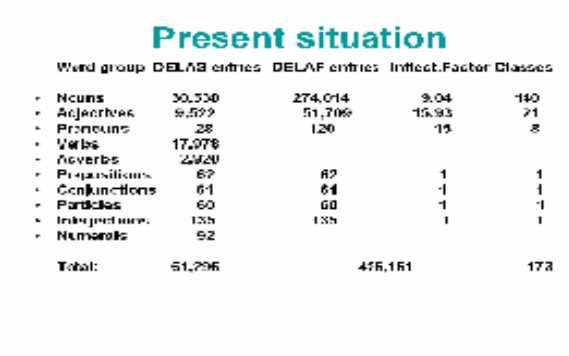

Figure 2: The total number of DELAS and DELAF entries (Petrovski, 2005)

Future activities are related to the compilation of the electronic dictionary of compound words (DELACF) and set of local grammars that can be used for disambiguation when a text is being processed. Furthermore, the Macedonian module should be adapted to the new version of the system called Nooj (Silberztein, 2005) which presents several differences when compared to Intex especially the organization and the compilation of the dictionaries. The basic feature of the Nooj dictionaries is the absence of the dictionary of inflected forms DELAF and the co-existence of both simple and compound words in a same dictionary. Some of the other main innovations in Nooj are: processing corpora rather than single texts, processing of more than 100 file formats which makes the system quite flexible and easy-for-use etc.

The development of a rich Macedonian module for Nooj will allow the linguists to use Nooj for processing of Macedonian corpora as well as a tool for extraction of terminology etc.

Figure 3. Example of text in Macedonian processed with INTEX (Petrovski, 2005)

3.2 Morphological dictionary

The second electronic dictionary is being constructed by the recently established LTRC (Language Technology Research Center). It is a morphological dictionary that associates each inflected form with a lemma and flectional information represented by different tags.

Various sources were used for extraction of words that were included in the initial database: lexicons, corpuses of texts retrieved from the Internet etc. Once the raw database of word forms was completed, the team had to extract the lemmas and to elaborate a methodology for generation of all inflected forms. This proved to be a difficult task since the Macedonian language is characterized by complex morphological system both on derivational and inflective level. The analysis and the generation of the inflected forms was done semi-automatically: around 50 inflectional paradigms for 10 word classes were developed and each lemma was assigned a code referring to the corresponding paradigm. The new expanded database was then manually corrected since many of the lemmas show inconsistencies with the corresponding inflectional rules.

At the same time the LTRC group elaborated a tag set which was used for the annotation of the word forms.

	CLASS
NN	noun
V	verb
ADJ	adjective
ADV	adverb
PREP	preposition
CONJ	conjunction
PRON	pronoun
NUM	numeral
PART	particle
INTERJ	interjection
ABBR	abbreviation
ADJPRON	adjective/pronouns

Figure 4: The word classes in the LTRC dictionary

Figure 4 shows the tags used for different word classes which is almost identical to the tags of the previous dictionary and of the Orwell's "1984" annotated corpus.

	SUBCLASS		**SUBCLASS**
VOC	vocative	ADJC	comparative adjective
OBL	oblique	ADJS	superlative adjective
DIM	diminutive	VADJ	verbal adjective
AUG	augmentative	PERS	personal pronoun
PEJ	pejorative	REL	relative pronoun
PR	present	POS	possessive pronoun
IM	imperfect	DEM	demonstrative pronoun
AO	aorist	IND	indefinite pronoun
IML	imperfect l	VADV	verbal adverb
AOL	aorist l	NUM	numerals

Figure 5: Subtypes in the LTRC dictionary

Still the tag set of the dictionary slightly differs from the MULTEXT-East notation system as far as the organization of the attributes and values is concerned presented in Figure 5. Beside this two columns that represent the type and the subtype of the word, there are several others used to insert information about the gender, number, article (three different types of articles added as suffixes to the nouns), case (if any form) and identification number. Currently the dictionary contains 1.535.668 generated word forms distributed as follows:

ABBR	249
ADJ	603863
ADJPRON	128
ADV	13938
CONJ	57
INTERJ	189
NN	407351
NUM	288
PART	56
PREP	63
PRON	413
V	509073
total	**1.535.668**

Figure 6: Distribution of the word classes

The high number of nouns can be explained with the relatively large database of proper names included in the dictionary. The high number of word forms which are adjectives is due to the fact that the comparative and the superlative are analytical.

The dictionary was tested on a half a million word corpus and managed to recognize 99.02% of the words.

4. Conclusions and future work

The corpora are intended as a source of data for linguistic research. They will help to capture all the meanings of a word, their frequency and the context in which they appear. The information regarding the relevance and frequency of each meaning can be incorporated in the lexicon of the Macedonian language. The reference corpus can also be used for more complex research, for detection of patterns of words and for the enlargement of the Macedonian morphological lexicon. The corpora, as well as the dictionaries will be used in future to build various NLP tools.

5. References

Istrazuvacki centar za arealna lingvistika (2005). Skopje: Makedonska Akademija na naukite i umetnostite

Mitrevski, G. Makedonski elektronski korpus: dizajn, implementacija, pristap. In *Predavanja na XXXVIII megunaroden seminar za nakedonski jazik, literatura i kultura.* Skopje: UKIM, Megunaroden seminar za makedonski jazik, literatura i kultura. In press.

Petrovski, A. (2005) *Macedonian DELAS- first results* laseldi.univ/fcomte.fr/document/colloque/nooj_2005/powerpoint/petrovski.ppt

Petrovski, Aleksandar. Za makedonskata kompjuterska leksikografija. In *Jazicnata politika. Informatikata i lingvistikata. Denovi posveteni na Blagoja Korubin, maj 2005*. Skopje: Institut za makedonski jazik Krste Petkov Misirkov. In press.

Silberztein, M. (2004) *Intex*

http://msh.univ-fcomte.fr/intex/downloads/Manual.pdf

Silberztein, M. (2005) *Nooj*

http://perso.wanadoo.fr/rosavram/NooJ%20Manual.pdf

Venovska-Antevska S. (2005). Makedonski jazicen korpus. Ideja, moznosti, realizacija. In *Predavanja na XXXVII seminar za makedonski jazik, literatura i kultura. Skopje: UKIM, Megunaroden seminar za makedonski jazik, literatura i kultura*. pp. 77-92.

Vojnovski, V., S. Dzeroski and T. Erjavec, (2005) Learning POS Tagging from a Tagged Macedonian Text Corpus. In *Proceedings of SIKDD 2005*, Ljubljana, 2005. In press.

Zdravkova, K., A. Ivanovska, S. Dzeroski and T. Erjavec, (2005) Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns. *In Proceedings of SIKDD 2005, Ljubljana, 2005*. In press.