5th
LREC 2020 Marseille
Workshop on Indian Language Data: Resources and Evaluation (WILD^RE)
rescheduled to 24th May 2020
Assamese . Bengali . Bodo . Dogri . English . Gujarati .
Hindi . Kannada . Kashmiri . Konkani . Maithili . Malayalam . Manipuri . Marathi . Nepali . Oriya . Punjabi . Sanskrit . Santhali . Sindhi . Tamil . Telugu . Urdu
(Organized under LREC2020, May 11-16, 2020)

Polish Lexicon-Grammar Development Methodology as an Example for Application to other Languages

Zygmunt Vetulani[1], Grażyna Vetulani[2]
[1,2] Adam Mickiewicz University in Poznań
[1] Faculty of Mathematics and Computer Science
[1] ul. Uniwersytetu Poznańskiego 4, 61-614, Poznań, Poland
[2] Faculty of Modern Languages and Literatures
[2] al. Niepodległości 4, 61-874, Poznań, Poland
vetulani@amu.edu.pl, gravet@amu.edu.pl

LREC 2020, Marseille, France
(on-line)

(Organized under LREC2020, May 11-16, 2020)

Polish Lexicon-Grammar Development Methodology as an Example for Application to other Languages

The full paper in the Wildre-5 proceedings:

https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/WILDRE-5book.pdf

The full presentation will be soon at:

https://drive.google.com/open?id=19vNUdKMXhkubuG33uTzdhM-PLx6UrNQc

LREC 2020, Marseille, France
(on-line)

Hierarchy

English is commonly considered as an absolute reference point for languages classification in terms of adaptation of their description to technological needs as well as in terms of richness of tools and language resources necessary for industry.

Below we locate the class of languages with reasonable language resources and tools that allow correct language-technological growth. Most European and several Indian languages are in this group. These may be considered to be "uper" or at least „middle-resourced".

Next go languages referred to as "less-resourced". In this group, we classify some minority languages from, by the way, technologically highly developed countries. Until recently the Polish language was categorized within this group.

At the very bottom of the hierarchy we find a significant number of languages for which there is no well defined (or realistic) needs to develop such technologies.

We address this paper to researchers working on **"middle" or "less-resourced"** Indo-European languages as **an offer of a long-term academic cooperation** in the field, within which we wish to share experience with our partners in the area under consideration.

From the paper abstract:

The reason of presenting some our works on lexicon-grammars within the Wildre workshop is the intention, among other, to take up the challenge thrown down in the CFP of Wildre, which is "to provide opportunity for researchers from India to collaborate with researchers from other parts of the world".

In the paper we present our methodology with the intention to propose it as a reference for creating lexicon-grammars. We share our long-term experience gained during research projects (past and on-going) concerning the description of Polish using this approach. This methodology, linking semantics and syntax, revealed useful for various IT applications. Among other, we address this paper to researchers working on "less" or "middle-resourced" Indo-European languages, as a proposal of a long-term academic cooperation in the field.

We believe that the confrontation of our lexicon-grammar methodology with other Indo-European languages, but also Non-Indo-European languages of India, Ugro-Finish or Turkic languages in Eurasia, will allow for better understanding of the level of generality of our approach and, last but not least, will create opportunities to intensify comparative studies.

Why lexicon-grammar?

Lexicon-grammars were a response to the needs of the language industry.

The language industry request was challenging :
to meet the needs of rigorous, exhaustive and easy to implement language models, as well as the needs of rigorous language descriptions.

The concept of **lexicon-grammar** answers to these needs: the main idea is to link an important amount of grammatical (syntactic and semantic) information directly to the respective words (verbs and other predicative words, i.e. words that open syntactic positions in a sentence).

Lexicon grammars were first systematically explored by Maurice Gross (Gross 1975, 1994), initially for French (fr. *lexique-grammaire*), then for other languages. Gross was also – to the best of our knowledge – the first to use the term lexicon-grammar.

In the linguistic tradition a crucial role in language description was typically given to dictionaries and grammars. Our memory goes back to Summerian-Akkadia dictionary tablets dated 2300 BC and first preserved Sanskrit grammars attributed to Yasdaka and Panini (6th century BC).
Until recently, dictionaries and grammars were used mainly for teaching and translation and were supposed to be interpreted by humans.

Nowadays being human-readable is not enough.

New concepts of organization of language description for better facing technological challenges emerged. Among them was the concept of **the lexicon-grammar**.

Our message addresses in particular two classes of languages:
- first, languages as Hindi or Sanskrit, with a rich linguistic tradition and preexisting language resources, for which the methods described in the paper will be easy to apply and beneficial.
- secondly, a multitude of languages spoken in the Indian subcontinent that do not have such a privileged starting position. In this case, in order to benefit from the methodology we present in the paper, an effort must first be done to complete the existing gaps, first of all concerning machine-readable grammars and dictionaries *("no royal road to the language techologies", adapted from Euclid's of reply to the king Ptolemy I, 367 BC - 282 BC).*

Participation in the two EU projects, CEGLEX – COPERNICUS 1032 (1995-1996) and GRAMLEX – COPERNICUS 621 (1995-1998), was an important first step towards the Lexicon Grammar for Polish and the real scale practical applications.

The further steps (described in the paper) were:

- continuation of earlier works on systems with language understanding competence where dictionaries were organised according the lexicon-grammar methodology (cf. POLINT-112-SMS),
- works on formal description of compouned verbs (verb-noun collocation)
- development of the WordNet like lexical ontology PolNet – Polish Wordnet; PolNet 1.0 (2011)
- extension of PolNet to a lexicon-grammar through integration of verb synsets; PolNet 3.0 (2016)

We consider this work as our contribution to bringing Polish out of the class of "less-resourced languages".

अनुगृहीतोऽस्मि

Thank you