Database of Multiword Expressions for Lithuanian (WG1)

Justina Mandravickaitė
Baltic Institute of Advanced Technology
Saulėtekio 15, Vilnius, Lithuania
justina@bpti.lt

Tomas Krilavičius
Baltic Institute of Advanced Technology
Saulėtekio 15, Vilnius, Lithuania
Vytautas Magnus University
Vileikos 8, Kaunas, Lithuania
t.krilavicius@bpti.lt

1 Introduction

We present the database of multiword expressions (MWE) for Lithuanian. In Lithuania most of the research was performed on collocations, therefore to start working on MWE a collection of MWEs is necessary. We believe, that this database will contribute to further explorations of MWE in Lithuanian.

2 Research and resources of MWE in Lithuanian

The research of MWE in Lithuanian for a long time was mainly commited to idioms, and since 2002 (development of Lithuanian national corpus, see http://tekstynas.vdu.lt) a major attention was given to collocations and morphological sequences. Formerwere mostly studied in (Marcinkeviciene, 2010), based on the predictability of selected collocations demonstrating that native speakers can successfully fill up missing words. It led to Lithuanian Dictionary of Noun Phrases (Rimkute, Bielinskiene *et al*, 2012), recurring noun phrases extracted from the corpus using Gravity counts (Daudaravicius, Marcinkeviciene, 2004) and verified manually. Collocations were analyzed w.r.t. translation (Volungeviciene, 2010, Kovalevskaite, Rimkute 2008), machine translation (Rimkute, Bielinskiene *et al*, 2008) and lexicography (Melnikiene and Jankauskaite, 2012).

Morphological sequences were analyzed in Kovalevskaite (2012), Kovalevskaite and Rimkute (2008), where they are defined as fixed sequences of words that cannot be analyzed semantically and syntactically, and have a single meaning as a part of speech function. Morphological sequences became of interest for linguists due to annotating Lithuanian corpus as certain words could not be annotated separately (Kovalevskaite, 2012).

3 Database of MWE for Lithuanian

We have chosen the most popular Lithuanian news portal delfi.lt. Corpus was formed from the headlines of the news articles, with an assumption that their purpose is to catch readers attention, by including phraseology or by the unusual choice of words/their sequences (not too unfamiliar).

Constructed corpus has >17 000 headings of news articles. Database was compiled of 2252 MWE examples of various length identified in the corpus. All of them belong to one of the three categories: traditional phraseological units (included in the Dictionary of Lithuanian Phraseology (2001)); metaphorical collocations ("frozen sequence of two elements (collocator and base), that have polysemous collocator, used in a figurative sense." (Volungeviciene, 2010); and other MWE (expressions which did not fit into the former two categories, mainly collocations common for media discourse). By category, there were identified 270 traditional phraseological units, 728 metaphorical collocations and 1254 other MWE. See Tables 1, 2 and 3 for examples.

Lithuanian	English
vienu šūviu nušauti du zuikius	to shoot two hares in one go, i.e. to do two things at once
varyti iš proto	to drive smb out of mind
užkurti pirtį	To kindle sauna/bathhouse, i.e. to haul/drag smb over the coals (to punish smb)
uždėti apynasrį	to put a bridle, i.e. to tame
padėti tašką	to put a dot, i.e. to finish
su žiburiu nerasti	do not find smth even with a lantern, i.e. to look for smth that is not there
šilta vietelė	little warm place, i.e. sinecure, well paid work that does not require efforts

Table **1**: Examples of MWE: Traditional phraseological units

Lithuanian	English
akmenėlis Kinijos bate	little stone in China's shoe, i.e. a small but very inconvenient problem for China
bado šmėkla	phantom of the famine, big famine or that the famine is expected
bėda viena nevaikšto	trouble never comes alone
grįžti su trenksmu	to come back with a bang, i.e. to come back noticeably
iš lūpų į lūpas	from lips to lips, i.e. word of mouth
medų su žmona kopinėti	to take honey from a hive with one's wife, i.e. to be on honeymoon
orų mergaitė	weather girl, i.e. weather newscaster

Table **2**: Examples of MWE: Metaphorical collocations

Lithuanian	English
30 sidabrinių	30 silvers, i.e. price of betrayal
Amerika ašaromis netiki	America (US) does not believe in tears, i.e. after the movie "Moscow does not believe in tears"
apetitas kyla bevalgant	appetite goes bigger while eating, i.e. the more you have, the more you want
audra stiklinėje	storm in a glass, i.e. big problem/ issue that in the end appears to be insignificant
blusų turgus	flea market
ant kilimėlio	on the little carpet, i.e. you are required (by your superior) to be reprimanded
gauti kaip šlapiu skuduru per veidą	as if to get a wet rag over the face, i.e. to get into unpleasant situation unexpectedly

Table **3**: Examples of MWE: Other MWE

Currently, the database is a text file (CSV, spreadsheet), containing MWE with its, frequency, type (Traditional phraseological unit, Metaphorical collocation or Other MWE), info about continuity and its location in the annotated corpus. All MWEs were identified manually by two annotators. Traditional phraseological units were identified using Dictionary of Lithuanian Phraseology (2001). It will be used for automatic identification of MWE in Lithuanian, e.g. using it for reference in combination with mwetoolkit[1].

4. Conclusions and Future Plans

We present a database of Lithuanian MWE – a new lexical resource for Lithuanian language as well as research on MWEs in general. At the moment it consists of MWEs and their variations, their frequencies in the corpus and information about their continuity. We plan to add lexical, syntactic and semantic information and make this database available publicly.

Acknowledgments

We thank Ugnė Civilkaitė and Laura Vilkaitė for comments. This research was partially funded by COST (Parseme, IC1207) and ESF (DADA, VP1-3.1-ŠMM-10-V-02-025).

References

E. Rimkutė, A. Bielinskienė, and J. Kovalevskaitė. 2012. Lietuvių kalbos daiktavardinių frazių žodynas. Vytauto Didžiojo universitetas.

R. Marcinkevičienė. 2010. Lietuvių kalbos kolokacijos.

V. Daudaravičius and R. Marcinkevičienė. 2004. Gravity counts for the boundaries of collocations. Int. Jrnl. of Corpus Linguistics, 9(2):321–348.

E. Rimkutė, A. Bielinskienė, and J. Kovalevskaitė. 2008. Kalbos pusfabrikačiai – sunkus maistas mašininio vertimo sistemos skrandžiui. Gimtoji kalba, 3:3–11.

J. Kovalevskaitė and E. Rimkutė. 2008. Morfologininių samplaikų struktūros ypatumai: kelių kalbų palyginimas.

J. Kovalevskaitė. 2012. Lietuvių kalbos samplaikos. Ph.D. thesis, Vytauto Didžiojo universitetas.

J. Paulauskas, ed. 2001. Frazeologijos žodynas. Lietuvių kalbos institutas.

S. Volungevičienė. 2010. Metaforinių kolokacių vertimo problemos. Kalbų studijos, (16):23–27

[1] http://mwetoolkit.sourceforge.net/PHITE.php?sitesig=MWE