JOLANTA KOVALEVSKAITĖ, ERIKA RIMKUTĖ,
JURGITA VAIČENONIENĖ

IDENTIFICATION OF LITHUANIAN ARBITRARY COLLOCATIONS

Summary

In this article, we describe the methodological approaches of arbitrary collocation (AC) identification developed within the framework of an ongoing project "Arbitrary Collocations of Lithuanian: Identification, Description and Usage (ARKA)". The object of this research is arbitrary collocations of the Lithuanian language the collocability of which is determined by lexical rather than semantic constraints. Such collocations do not directly reflect the non-linguistic reality and the relation between the collocates is arbitrary: although there may be several close synonyms, a particular one is preferred in a certain word combination (e.g., *suprasti užuominą* vs. **suvokti užuominą*).

The source of the research was the Database of Lithuanian Multiword Expressions encompassing over 12.000 collocations. The structural composition of the identified 2400 arbitrary collocations was as follows: adjective (participle)+noun, verb+noun, and noun+noun.

Arbitrary collocations were determined by combining the manual linguistic analysis and semi-automatic methods of computational linguistics. The manual method included two major AC identification criteria: lexical restrictedness and (or) meaning transfer. Lexical restrictedness was measured using two tests: the synonym substitution of pre-modifier and (or) semantic field comparison of the head noun. Moreover, lexical restrictedness was also assessed using a semi-automatic approach by analyzing synonym pairs in the automatically generated vector strings. The semi-automatic approach consisted of three stages: (1) the automatic generation of vector strings with potential synonyms; (2) the manual vector string editing to reduce the noise, and (3) the semi-automatic process during which the collocations were compared to particular synonym pairs in vector strings and approved or not approved by linguists as arbitrary.

Approximately half of ACs were detected by using the manual method based on two AC identification criteria, whereas about one-third of ACs were identified using the semi-automated method based on one criterion. On the one hand, the semi-automatic method can reduce the subjectivity in data evaluation and simplify the AC identification procedure. On the other hand, as arbitrariness is determined on the basis of the available synonyms in the vector string, some ACs may not be identified because of the insufficient scope of the generated synonyms. Also, the automatically generated synonyms need to be manually analyzed to eliminate semantically unrelated words. Despite these limitations, the method is suitable for processing large amounts of the data. It would be difficult to manually assess the contextual similarity of such a large number of words.

About one fifth of ACs were identified using a combination of both methods. These results suggest that different methodological approaches allowed the researchers to detect more arbitrary collocations. It is maintained that taking into consideration the mentioned advantages and disadvantages, the best results in the Lithuanian arbitrary collocation identification are acquired when combining both manual and semi-automatic methods.

KEYWORDS: arbitrary collocations, Lithuanian, lexical restrictedness, vector space model, semi-automatic approach, linguistic manual approach.

JOLANTA KOVALEVSKAITĖ, ERIKA RIMKUTĖ,
JURGITA VAIČENONIENĖ
Vytauto Didžiojo universitetas
Kristijono Donelaičio g. 58, 44248 Kaunas
jolanta.kovalevskaite@vdu.lt
erika.rimkute@vdu.lt
jurgita.vaicenoniene@vdu.lt