Adapting lexicographic tools to automated phraseme recognition in a machine learning system

On the example of Russian verb-noun phrases

Summary

This monograph is methodology-oriented. Its main body includes a series of experiments which all lead to the automation of recognition of particular multi-word structures and their classification as either phrasemes (F) or non-phrasemes (N-F). Thanks to the technical solution to the categorisation of the phrases under investigation, coupled with manual tagging of these structures interpreted within a special phrasematic key, it was possible to improve the *machine learning* system (ML). This multimodal learning system used experience (i.e. manually processed material) to automatically identify the so-called *potential* phrasemes in a randomly selected sample of a Russian-language text. Such potential phrasemes identified by artificial intelligence (AI) have been labelled *bingo*. Itemisation of the *bingo* unit was the main aim of this research.

The research unit proved to be a verb-noun phrase (VNP) of a particular morpho-syntactic structure, pre-determined choice and range of grammatical structures (a cluster of grammemes) of each component. Two types of research units are identified in this monograph, i.e. input and output ones. The input units are those verb-noun phrases which comprise the *potential* phrasemes VNP-PP$_{\text{Source Material R}}$ (e.g. *давить на психику*, *ходить на работу*, *забежать на минутку*), which were selected with a view to using the ML (together with the distribution context) as the source material comprising the experience of the system. The VNP-PP$_{\text{Source Material R}}$ units in the Source were labelled with the phraseme qualifier. The qualifiers are placed on a phraseme scale, which in turn was used to create the phraseme key composed of qualifiers referring the extreme ends of the scale, i.e. from F to N-F. On the other hand, VNP-PP$_{\text{output-bingo}}$ (e.g. *баллотироваться на должность*, *вскочить на подножку*, *вскочить на грудь*) comprises the output units that the ML identified in the sample text and assigned them the phraseme characteristics in the form of the phraseme qualifier, following the pattern of the qualifiers in Source R. It has to be highlighted that one of the more important subtypes of VNP-PP$_{\text{output-bingo}}$ comprises unique units, i.e. those that do not have their lexical duplicates in the VNP-PP$_{\text{Source Material R}}$. In other words, the ML is capable of identifying any VNP that qualifies as a phraseme in a random text, or discriminate between F and N-F. As a consequence, it

can be used, together with precise morpho-syntactic characteristics, to parametrisize an idiolect, which in practice makes it possible to verify the authorship of the text from the phrasematic and morpho-syntactic point of view.

The lexicographic tool described in this monograph is a practical realisation of the idea of the automated recognition of multiword structures in a phraseme key (ARMSPK). The research tool within the linguistic basis (BL) is labelled here as the ARMSPK System (i.e. SARMSPK), supported with the linguistic, and, to a lesser degree, also an IT-related explanation of the structure of analytic operations, i.e. research tools such as techniques of excerption, selection and verification.

The aim of the series of the experiment phases is to provide the learning system with the kind of experience-source that would enable it to identify phrasemes without the user's intervention. Moreover, this kind of method is aimed at automated indentification by the ML of the potential translation language pairs (both those that are phrasemes and those which are not) in the parallel Russian – Polish language corpus. In order to be able to "feed" such source language material into the system, it had to be prepared from the theoretical point of view (i.e. take up a particular methodological stance, delineate the notions' boundaries), as well as practical (i.e. tag the words appropriately so that the SARMSPK would be able to identify them). Providing the system with such appropriately prepared source material (Source R) was the main objective of this piece of research. Therefore the focus was on the methodology of creating and providing the source (experience) to the system in the form of randomised Source R (i.e. divided into pre-determined subgroups, while retaining its consistent form throughout all phases of the research procedure, such as preadaptation, adaptation and postadaptation). The source material fed into the ML with the aim of automated recognition of phrasemes in the given text can be characterised as representative: 1) it comprises a fairly big number of units of an identical morphological range. Moreover, the spectrum of grammeme clusters is specified for each of the elements of such a multiword unit; 2) it has been verified from the point of view of the phraseme characteristics of such units; when creating the source material, the focus was on diminishing the subjective judgement of a multiword unit (from the point of view of the phraseme characteristics) thanks to several parameterised phases of verification. In other words, the research tool describes the linguistic base, i.e. the parameters identified in theory have been transposed into the research material (Source R). The research material has been appropriately adapted for the SARMSPK system to recognise. The IT base (BI) comprises an appropriate absorption and processing, i.e. "understanding" of the Source R containing numerous VNP-PP of a particular morpho-syntactic type that is either F or N-F. As for the IT base, several research aspects have been analysed (cf. Subchapters I.2.1–I.2.7 containing the word *discussion*), which focus on such notions as phraseme and *reproducibility*.

Adaptation, i.e. the absorption and adjustment of the linguistic base (BL) to the IT base (BI), involves providing the SARMSPK with the language source material (i.e. Source R). Such adaptation requires a well-defined research approach. The research process comprises three phases: preadaptation, adaptation, postadaptation. It is reflected in the contents of this monograph, which is composed of four chapters (one theoretical and three practical):

I. *BL Background: research unit and Source R.* This is the theoretical part, i.e. the description of the BL background, introductory characteristics of F and N-F.

II. *SARMSPK Preadaptation Methodology: Source R retrieval phases.* This is the pre-adaptation part (also referred to in the monograph as proto-adaptation), i.e. VNP excerption, verification and qualification on the phraseme scale; retrieving the pre-adaptation trial source material (Source R) for the SARMSPK.

III. *SARMSPK Adaptation Methodology.* This part contains the description of the language material conversion in line with the IT base requirements; feeding data into the SARMSPK.

IV. *SARMSPK post-adaptation Methodology: return to the linguistic formalism.* This part of the monograph deals with post-adaptation and contains a description of the analysis the SARMSPK results obtained, identifying the *bingo* units.

Apart from the Introduction and the four chapters, the monograph contains a Conclusion, Bibliography and a verification resource list (containing mainly dictionaries, e.g. the *Registration Fact Megalist (RFM)*, a terminology list together with their description entitled *Terminology and technical description*, a list of tables, figures and charts, i.e. *Tables and Illustrations*). A compact disc is also an integral part of the monograph, as it contains both the source material data and SARMSPK results. An additional chapter, entitled *Appendix. CD-ROM*, contains a "step-by-step" description of the way the data was obtained.

In light of the main objective of this monograph (obtaining the bingo unit), it is worth concluding that from the point of view of the methodology used in this research, it seems to be optimal, as it appears to coincide with the main objectives, i.e. it is methodologically self-reliant and interchangeable, verifiable, but most importantly, accurate in terms of its methodological techniques and practical nature, as it leads us to theoretical pre-adaptation assumptions (referring mainly to the reproducibility and autonomy of potential phrasemes) supporting the preparation of the Source R for being fed into the SARMSPK. The results obtained serve as evidence that the methodology described here can be utilised to analyse various other morpho-syntactive structures (other than VNP) characterised by the idiomatic potential.

Translated by Rob Pagett