NEW BULGARIAN UNIVERSITY

DEPARTMENT OF FOREIGN LANGUAGES AND CULTURES

ENGLISH LANGUAGE

REVIEW

BY ASSOC. PROF. **VALENTINA IVANOVA GEORGIEVA**, PhD
"G. S. RAKOVSKI" NATIONAL DEFENCE COLLEGE - SOFIA
Associate Professor in the Professional Field 2.1. Philology,
Subject Area: General and Comparative Linguistics

OF A DISSERTATION
FOR OBTAINING THE DEGREE OF DOCTOR OF SCIENCE

Professional Field: 2.1. Philology (Subject Area: Translation theory and practice)
by the candidate Velislava Stoykova, PhD
TOPIC: "Machine Translation Approaches to Bulgarian Language"

Velislava Stoykova's D.Sc. dissertation sets as its main **research goal** to analyse the approaches of formal language theories to be allied in the computer processing of Bulgarian language data for machine translation (MT), as well as their limitations. This goal is formulated in the executive summary of the thesis, written in Bulgarian (p. 2), while the foreword of the thesis (written in English) states that "The book presents a Contrastive Analysis study of MT approaches to the Bulgarian language for cross-lingual translation. The study includes the analysis of ***several*** (italics mine) Bulgarian language MT application software using different approaches and techniques to process linguistic data. The results of the software analysis are presented to outline the limitations in scope and performance to improve the quality of Bulgarian language MT." (p. vii) The study aims to achieve its goal by solving some **practical tasks,** i.e. to present several Machine Translation software applications for the Bulgarian language by presenting parts of their source code and comparing them with those used for the target languages in translation; analysing the scope of these applications and the quality of the translation they produce (p. 2). The goals and tasks of the reseach are slighly differently articulated in the dissertation and in the executive summary of the dissertation, so it would be of higher acadmic value if they were precisely formulated and an affirmation of the accomplishment of the research tasks were presented in the conclusion. It should be noted that the executive summary of the dissertation in Bulgarian is more precise as it says that the contrastive analysis (CA) is applied in order to investigate ***three*** approaches to MT (p. 2 of the Summary).

The clear atatement of the **importance of the researched problem** can also be found only in the executive summary of the dissertation (p. 2 of the Summary): „There is no detailed research in the academic literature on the boundaries and the efficiency of using formal language theories for Bulgarian language."

The D.Sc. **dissertation content** is 95 pages, which iclude: the theoretical part; the bibliography of 5 pages and 105 references; 4 pages with Appendices to the different chapters of the book, and 2 pages with Glossary. The content of the dissertation is structured in five chapters: Introduction, Machine Translation Using the DATR Language for Lexical Knowledge Representation, Machine Translation Using the Universal Networking Language, Machine Translation Using the Sketch Engine, and Conclusion. Thus, the three main chapters of the research elaborate on the contrastive analysis of these three techniquess of MT for the Bulgarian language, although, as the author states, „In Bulgarian theory and practice of translation, there are authors presenting all the above approaches" (p. 5).

The **research methodology** is presented in **the Introduction** and is said to be based on the contrastive analysis (CA) since, as V. Stoykova states, „the CA distinguishes and describes only similarity (on the basis of semantic equivalence analysis) from a cross-lingual perspective" (p. 4). She adds, „CA starts where the general linquistic ends" (p. 4), which is a controversial claim, because in the next paragraph the author says, „CA at all levels of description (phonetics and phonology, morphology, and syntax) is oriented toward language structure and is centered on the form" (p. 4) which is also a subject of study of general linquistics. In addition, the author introduces and explains the concept ot *tertium comparationis* as a key concept for contrastive analysis (p. 3) because „TC is based and operates on the use of semantic equivalency to identify or to characterize as similar the compared linguistic phenomena" (p. 4). Thus, the chosen methodology contributes to the reliability and representativeness of the research and is a solid basis for conducting the analysis and drawing some conclusions.

After presenting the key differences between human translation and machine translation (MT), the researcher narrows her analysis to the two types of MT, namely Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). We would suggest presenting some other approaches to MT as well, e.g. Neural MT, Hybrid MT, Example-based MT and/or other types, comparing them and justifying the choice of the researched two types which are analysed in detailes.

Chapter 2 presents the working mechanism of DATR Language for Lexical Knowledge Representation. One of the contributions of the dissertation is the analysis of linguistic and computational approaches to inflectional morphology (2.3) and the suggestion for the application of the "mixed" approach for MT as the most appropriate since both stems and suffixes are variables. In 2.4 the author presents the grammar inflectional rules in a semantic network and graphically presents the Bulgarian nominal inflectional type hierarchy in DATR, while in 2.5. Russian nominal inflectional morphology in DATR is presented. The important conclusion which is drawn as a result of the semantic analysis of the word 'zakon' is presented only in the Summary: although homonymous in Bulgarian and Russian languages, „DART is a reliable FLT for multi-lingual MT between Slavic languages (p. 9 Summary)"; the computer programme cannot mix the 2 words because they are encoded with their specific grammar (inflectional rules), and the interpretation uses declensional classes and the features of the case, number, and animacy. A conclusion similar to the one formulated in the author's executive summary is made after the presentation of the Polish mominal inflectional morphology in DATR, which is more schematic than that for Bulgarian and Russian, „„... related languages can be formally presented by using similar ideas and techniques for encoding because they share similar grammatical features." (p. 20)

Another contribution of this chapter is the DATR software code for clarifying the grammatical categories of possessive pronouns and the reflexive possessive pronoun in Bulgarian.

Chapter 3 is dedicated to the analysis of the Universal Networking Language (UNL) for Machine Translation. The author justifies its 'superiority' over DATR by presenting its characteristics in 3.1. and underlying the fact in 3.2.2. that „The UNL application represents online web-based intelligent information and a knowledge management system which allows the use of different types of semantic relations for linguistic data organization." (p. 32). Since UNL possesses better capabilities for MT, the encoding features for the lexical database in Bulgarian, English, Russian and Slovak UNL used for MT are described in the chapter.

The contribution of this part of the research is the comparison of the two types of encoding of the Bulgarian nominal inflectional morphology (DATR and UNL) and the reveal of their limitations of the particular formalism.

In 3.3. V. Stoykova analyses the formal representation of possessive pronouns in Bulgarian, English and Russian in the frameworks of the UNL. After theoretical explanations of the UNL types of transformation inflectional rules, the process of encoding is exemplified with the UNL Dictionary for the English pronouns 'mine' and 'myself', the Bulgarian 'moj', and the UNL lexical entry for the Russian pronoun 'moj' The conclusion drawn by the author is that „The UNL pronouns presentation scheme is capable of offering adequate encoding not only for related languages (like Bulgarian and Russian) but also for non-related languages (like English and Bulgarian or English and Russian) and is successfully used for MT. The application is open for further improvement and development by introducing additional grammatical rules and enlarging the database." (p.37).

Section 3.4 is dedicated to anaysing the UNL-based MT between Bulgarian and Slovak. The author's contribution is the presentation of the grammatical features of both Bulgarian and Slovak by employing the UNL encoding as an inter-lingua presentation on the base of the related translations of the book 'The Little Prince'. The encoding rules are exemplified with the lexical entries for the word 'dobyr' in Bulgarian and 'dobrý' in Slovak. The conclusion is that the „UNL applications of Bulgarian and Slovak encoding for MT purposes use in full scale the UNL mechanism to present adequately grammatical and lexical knowledge for both languages" (p. 42).

The final section 3.5. is focused on measuring and improving the accuracy of the UNL-based MT of the possessive pronouns in Bulgarian and English. The author claims that „the UNL web-based platform makes possible the statistically-based measurement of the accuracy of the translated texts by the estimation of precision and recall". Her contribution is the development of transformation syntactic rules to differentiate possessive pronouns as modifiers from those as specifiers which have significantly improved the accuracy of the translation (p. 48).

Chapter 4 presents the Sketch Engine (SE) as a search engine and its capabilities for MT. The introduction of SE reveals the main principle of SE which is based on the assumption that if two words are statistically similar, then they are semantically related (similar principle applies to the MT of Google Translator).

Section 4.1.2. describes Sketch Engine statistical properties as *tertium comparationis.* Thanks to the SE properties, a contrastive analysis and MT of ecological terms in Bulgarian and Slovak is conducted by employing the bilingual parallel text corpora EUROPARL 7 search using the SE statistical scoring. The possible relations between the Bulgarian and the Slovak translation of ecological terms equivalence are studied and analyzed. The conclusion of the researcher is that the corpus is a reliable source for translating and analyzing the linguistic structure of more popular ecological terms for both languages, as well as to investigate the similarity in their bilingual linguistic translations, although ecological concepts for Bulgarian and Slovak are not always symmetrical, e.g. 'биологично разнообразие' (BG) – 'biodiversity' (SK).

What follows in the dissertation is Sketch Engine-based MT of time expressions between Bulgarian and Slovak using the existing Bulgarian-Slovak electronic text corpora. The SE techniques of keywords search for concordances and collocations generation from the parallel sentence-aligned bilingual electronic text corpora as TC and inter-lingua is used to compare the text corpora semantic content as well as the generated time expressions translations. As an example, the correlation of the Slovak word 'obdobie' (period, term) and its possible translation(s) into the Bulgarian language by the generation of parallel bilingual concordances of that keyword from the Bulgarian-Slovak Corpus is presented. The examples suggest that statistical approach of the SE is reliable for the correct translation of words because it accounts for the word (keyword) semantic properties and can differentiate among very subtle word semantic features to maintain good translation equivalence for correct translation.

In 4.4. SE is analysed for its MT properties for translation between Bulgarian and Slovak languages and the search for words with common collocations reveals hidden semantic relations which can be visualized with different colours. SE makes it possible to search from the bilingual parallel corpora of texts.The conclusion is that the proposed research presents a statistical approach to MT for Slovak to Bulgarian language translation. The main advantage, as undelined by the author, is that SE is very fast and it does not require FGs for the source and the target languages, only sufficient bilingual specialised corpora.

Section 4.5. presents another contribution of the research with the design of The Bulgarian Mathematical Wikipedia Corpus (MathWikiBG) and the Serbian Mathematical Wikipedia Corpus (MathWikiSR) are the first Bulgarian–Serbian / Serbian–Bulgarian comparable electronic text corpora using SE as *tertium comparationis*. Algorithmic steps of keyword search in the first comparable web-generated corpora in Bulgarian and Serbian are employed. This is the only place where the author states that one of „The key contribution here is that we have expanded the keyword search with language-specific tagging (allowing for the processing of different parts-of-speech, inflection, etc.) for both Bulgarian and Serbian." (p. 75).

The D.Sc. dissertation's scientific contributions have been underliend in the course of the review, but it is worth mentioning that the research not only presents and summarises the properties of the three MT techniques applied to the Bulgarian language (i.e. DATR, UNL and SE), but also compares and analyses their pros and limitations on the basis of sufficient examples. We cannot but agree with the candidate's self-assessment of the main contributions formulated in the executive summary on p. 34. It is important to stress that V. Stoykova designs a corpus of specialised texts in Mathematics in Bulgarian *MathWikiBul* which is used for statistical MT between Bulgarian and Serbian Languages.

The **references** to theoretical studies demonstrate deep comprehension of the conceptual framework on issues that are necessary to achieve the dissertation's goal, but some of them are quite outdated (e.g. from 1989) while the technological solutions to various research problems are updated rapidly.

The candidate presents 12 publications that complement the dissertation.

Recommendations and suggestions:

- I would encourage Dr Stoykova to emphasize more clearly her contribution to the academic research, instead of using impersonal sentences.
- The specialised academic text will be more easily comprehended if there was not such an overuse of abbreviations which hamper smooth assimilation of the arguments.
- In some places there are references to sources that require additional search of an article which is not available online, e.g. „We use TC as it is defined in [37]", „Our approach to the inflectional morphology of possessive pronouns is based on the encoding already published in [63]" (p. 21), „The detailed DATR account of the Bulgarian nominal inflectional morphology has already been published in [59]" (p. 30).

The executive summary of the dissertation summarises the content of the dissertation but with some of the formulations it is more precise and explicit.

Based on the dissertation research contributions and the positive qualities highlighted above, it can be concluded that the dissertation meets the requirements for obtaining the D.Sc. degree. Therefore, I recommend that the Scientific Jury assess the dissertation on the topic of “Machine Translation Approaches to Bulgarian Language” positively and award the candidate Velislava Stoykova, PhD, the degree „Doctor of Science” in the Professional Field 2.1. Philology (Subject Area: General and Comparative Linguistics).

Signature:......................

/Assoc. Prof. Valentina Georgieva, PhD/

12.09.2022

Sofia