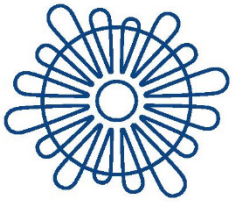

Sveučilište u Zadru
Universitas Studiorum
Jadertina | 1396 | 2002 |

THE 17th NOOJ INTERNATIONAL CONFERENCE 2023

Book of Abstracts

Zadar, May 31 – June 2, 2023

Editors

Linda Mijić
Department of Classical Philology
University of Zadar
Zadar, Croatia

Anita Bartulović
Department of Classical Philology
University of Zadar
Zadar, Croatia

Marijana Tomić
Department of Information Sciences
University of Zadar
Zadar, Croatia

Laura Grzunov
Department of Information Sciences
University of Zadar
Zadar, Croatia

Kristina Kocijan
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia

Max Silberztein
Université de Franche-Comté
Besançon, France

Publisher

University of Zadar
For the Publisher: Dijana Vican, rector

ISBN 978-953-331-421-1

Automatic Disambiguation of the Belarusian-Russian Legal Parallel Corpus in NooJ

Valery Varanovich
Mikita Suprunchuk
Yauheniya Zianouka
Yuras Hetsevich
Nastassia Yarash

United Institute of Informatics Problems
Minsk, the Republic of Belarus
ssrlab221@gmail.com

Abstract

Homonymy is still the central problem of automatic text processing at the lexical level. Automatic (rarely semi-automatic) solving of lexical ambiguity was first formulated within the field of science and technology related to the creation of machine translation systems. To date, this is a critical problem of improving the quality of systems for various branches of computational linguistics (Varanovich, 2009).

There are two dominant classes of ambiguity resolution mechanisms (Bouarroudj et al., 2022):

(1) Automatic one, implying a fully computerised solution to this problem.

(2) Interactive (dialogic, semi-automatic) one, supposing a joint solution by a person and a computer.

It means that the user has a set of alternatives from which he should choose one option.

We are conducting work on creating a legal texts corpus in Belarusian and Russian, which is used in various software products (speech synthesis, machine translation, spell checker, etc.). A trilingual Belarusian-Russian-English dictionary of legal terms was created (Hetsevich et al., 2021) within the project. Also, in the process of forming the corpus, contextual dictionaries are compiled, which can become the basis for high-priority dictionaries in the NooJ system. The dictionaries reflect contexts that indicate the preferred translation of a certain term from Russian into Belarusian (Barabash, 2015).

We are planning to create high-priority dictionaries for each of the 26 law codes of Belarus. It is assumed that a lot of diagnostic contexts will be the same in different codes (*настоящий* 'this' = *гэты* 'this' (code), not *настоящий* 'this' = *сапраўдны* 'real' (code)), but we hypothesize that in some cases the contexts for the same values will be different, and it is also possible that in different codes, one polysemous word (or a homonym) will have different meanings.

Here are some words with several meanings which were found during our work: данный – ‘1) гэты, this; 2) дадзены, given’; отпуск – ‘1) выдача (тавараў), issuance, supply; 2) адпачынак (даць), holiday, vacation’. These meanings (translations) could be chosen according to their neighbor words. As NooJ is an effective tool for solving word ambiguity, we plan to use it for compiling syntactic grammars. They will search for homonyms by analyzing the context (the sequence of words) and form a list of various lexical units for different domains. This will assist in identifying terms in different thematic domains, which is very important for compiling special vocabularies for indicated fields.

Key words

Disambiguation, homonym, automatic text processing, dictionary, legal texts corpus

References

[1] Barabash, O. V. (2015) Razgranichieniie omonimii i polisiemii juridichieskikh tierminov (= Барабаш О. В., Разграничение омонимии и полисемии юридических терминов). Rhema. Рема, 2, 39–51. Available at: https://cyberleninka.ru/article/n/razgranichenie-omonimii-i-polisemii-yuridicheskih-terminov.

[2] Bouarroudj, W., Boufaida, Z., Bellatreche, L. (2022) Named Entity Disambiguation in Short Texts over Knowledge Graphs. Knowledge and Information Systems, 64(2), 325–351. Available at: https://doi.org/10.1007/s10115-021-01642-9.

[3] Hetsevich, Y. et al. (2021) Creation of a Legal Domain Corpus for the Belarusian NooJ Module: Texts, Dictionaries, Grammars. In: Bigey, M. et. al. (eds.) 15th International Conference NooJ 2021: Book of Abstracts, 36–37. Available at: http://www.nooj-association.org/nooj2021/Abstracts.pdf.

[4] Silberztein, M. (2003–) NooJ Manual. Available at: http://www.nooj-association.org.

[5] Varanovich, V. V. (2009) Slovar lieksichieskikh valientnostiej v sistiemie russko-bielorusskogo mashinnogo pierievoda (= Воронович, В. В. Словарь лексических валентностей в системе русско-белорусского машинного перевода). In: Третьи чтения, посвященные памяти В. А. Карпова: сб. науч. ст., 108–111. БГУ, Минск. Available at: http://elib.bsu.by/handle/123456789/10645.