Composing Theoretical and Computational Linguistic Problems

Ivan Derzhanski
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Abstract

This tutorial will present the genre of the self-contained linguistic problem as a form of teaching facts about languages and concepts from linguistics, as well as the major techniques of composing problems and problem sets for linguistic contests, both classical and computational, but with a special focus on the latter.

It is intended for all who are interested in linguistic problems, whether because of current or envisaged participation in organising a contest, or interest in trying one's hand as a problem author, or involvement in teaching in any form, or plain curiosity.

Keywords: computational linguistic problems, linguistic problems, linguistic olympiads.

1 General description

For more than six decades, linguistics olympiads – contests for secondary-school students in solving linguistic problems – have served as a means of acquainting such students, but also the general public, with the science of language and its applications, thus filling a gap in school curricula from which this domain of human intellectual pursuit is typically absent.[1] Accordingly, the idea is that the problems should be self-contained — primarily inductive rather than deductive, not testing previously obtained knowledge but inviting the discovery of rules on the basis of unseen data, using only one's analytical skills and general culture in addition to such fundamental notions about language as are introduced in native and foreign language classes.

The contests were inspired by and modelled on mathematics olympiads, mathematicians cooperated with linguists in their design and implementation, and the word 'mathematics' or 'mathematical' was (and still is) often present in the title; but it was understood that what mathematics provided was the general method, the idea of abstract logical reasoning, not the content. The subfield of linguistics to which most linguistic problems can be said to be relevant is unquestionably typology.

Notwithstanding, since the earliest instalments of the first regular linguistic contest in history, the Moscow Traditional Olympiad in Linguistics and Mathematics (established in 1965), concepts and topics from computational linguistics have been present in it, in anticipation of the explosive development of language technologies that was destined to come shortly. Unsurprisingly, in recent years such problems have been appearing more frequently at 'classical' linguistic olympiads, and several contests have been launched which explicitly name computational linguistics as their primary or only subject, such as the North American Computational Linguistics Open Competition (NACLO, est. 2007) and the Australian Computational and Linguistics Olympiad (OzCLO, est. 2008), or the Contest in Computational Linguistics offered by the Institute for Bulgarian Language, Bulgarian Academy of Sciences. As the olympiads grow in numbers and size, the number of problem authors grows as well — but so does the need for problems.

The proposed tutorial, envisaged as **introductory**, will present the art – that is, the major techniques – of composing problems and problem sets for linguistic contests, both classical and computational, but with a special focus on the latter.

The **target audience** are primarily those who are engaged or interested in organising a contest in linguistics or computational linguistics, but also linguists, and especially computational linguists, both professionals and (postgraduate) students, who are involved (or considering being involved) in teaching at any level and would like to broaden the range of methods that they use. There is **no** prerequisite background.

[1] Linguistic problems have also been found highly useful in other educational activities, notably in university teaching.

The total duration will be **four** academic (forty-five-minute) hours. The tutorial will cover:

- an overview of the genre of the linguistic problem and the history of linguistic contests and olympiads (1 hour);
- a presentation of the process of composing (and solving) linguistic problems of the 'classical' (theoretical, typological) type (2 hours);
- a discussion of problems featuring concepts from computational linguistics as a subgenre of linguistic problems (1 hour).

The tutorial has **no** required preliminary reading. The exposition, however, will refer to published literature on linguistic problems as a genre (e. g., Зализняк, 1963; Алпатов et al., 1973; Журинский, 1993; Derzhanski, 2009); on linguistic contests and other extracurricular activities in linguistics, both in general (e. g., Derzhanski and Payne, 2010) and in individual countries (e. g., Derzhanski, 2007a; Hudson and Sheldon, 2013; Verhoeven, 2016; Lopuhaä-Zwakenberg, 2018); on the relationship between (theoretical) linguistics and mathematics as seen in linguistic problems (e. g., Derzhanski, 2007b; Derzhanski and Veneva, 2018); on the use of electronic language resources in developing linguistic problems (e. g., Iomdin et al., 2013; Derzhanski and Dekova, 2014); and on problems and contests oriented specifically towards computational linguistics (e. g., Littell et al., 2013; Estival et al., 2013, 2014).

2 The instructor

Ivan Derzhanski is an associate professor at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. His research interests are in (mostly computer- and corpus-aided) contrastive linguistics and typology, corpus linguistics, translation theory and practice and the methodology of linguistics. He teaches courses in Python programming for linguists, computational linguistics and natural language processing in undergraduate and postgraduate programmes at the University of Sofia and the New Bulgarian University. In Bulgaria he has been playing the central part in providing scientific support to the Olympiad and National Contest in Linguistics and to linguistic seminars for secondary-school students, putting together problem sets for the former and lecture programmes for the latter, continuously since 1996. He is a co-founder of the International Linguistics Olympiad (IOL, est. 2003) and has served on the problem committee and jury of all its instalments and chaired them thrice. Also he is a member of the team organising the national Contest in Computational Linguistics, centred at the Institute for Bulgarian Language, Bulgarian Academy of Sciences. He has authored and co-authored more than 120 linguistic problems.

3 Further details

An audience of about 20 people can be expected; since the tutorial has not been offered before, this (admittedly rough) estimate is based on the overall attendance of earlier instalments of the conference.

There are **no** special requirements for technical equipment.

The tutorial will be accompanied by handouts and a presentation. The materials can be made publicly available on the CLIB 2026 website.

References

Ivan Derzhanski. 2007a. Extracurricular activities in linguistics for secondary school students. In *Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia*, pages 125–128, Sofia, Bulgaria. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences.

Ivan Derzhanski. 2007b. Mathematics in linguistic problems. In *Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia*, pages 49–52, Sofia, Bulgaria. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences.

Ivan Derzhanski. 2009. *Linguistic Magic and Mystery*. Union of Bulgarian Mathematicians, Sofia.

Ivan Derzhanski and Rositsa Dekova. 2014. Electronic language resources in teaching mathematical linguistics. In *Proceedings of the First International Conference on Computational Linguistics in Bulgaria (CLIB 2014)*, pages 1–5, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Ivan Derzhanski and Thomas Payne. 2010. The Linguistic Olympiads: academic competitions in linguistics for secondary school students. In Kristin Denham and Anne Lobeck, editors, *Linguistics at School. Language awareness in primary and secondary education*, page 213–226. Cambridge University Press, Cambridge, UK.

Ivan Derzhanski and Milena Veneva. 2018. Linguistic problems on number names. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pages 169–176,

Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Dominique Estival, Cathy Bow, John Henderson, Barbara Kelly, Mary Laughren, Elisabeth Mayer, Diego Mollá, Colette Mrowa-Hopkins, Rachel Nordlinger, Verna Rieschild, Andrea C. Schalley, Alexander W. Stanley, and Jill Vaughan. 2014. Australia loves language puzzles: The Australian Computational and Linguistics Olympiad (OzCLO). *Language and Linguistics Compass*, 8(12):659–670.

Dominique Estival, John Henderson, Mary Laughren, Diego Mollá, Cathy Bow, Rachel Nordlinger, Verna Rieschild, Andrea C. Schalley, Alexander W. Stanley, and Colette Mrowa-Hopkins. 2013. Learning from OzCLO, the Australian computational and linguistics olympiad. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 35–41, Sofia, Bulgaria. Association for Computational Linguistics.

Richard A. Hudson and Neil Sheldon. 2013. Linguistics at School: The UK Linguistics Olympiad. *Language and Linguistics Compass*, 7(2):91–104.

Boris Iomdin, Alexander Piperski, and Anton Somin. 2013. Linguistic problems based on text corpora. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 9–17, Sofia, Bulgaria. Association for Computational Linguistics.

Patrick Littell, Lori Levin, Jason Eisner, and Dragomir Radev. 2013. Introducing computational concepts in a linguistics olympiad. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 18–26, Sofia, Bulgaria. Association for Computational Linguistics.

Milan A. Lopuhaä-Zwakenberg. 2018. De taalkunde olympiade. *Nieuw Archief voor Wiskunde*, 19(4):295–299.

Babette Verhoeven. 2016. The pleasures of a good problem: writing puzzles for the linguistics olympiads. https://doi.org/10.13140/RG.2.2.16297.85602.

В. М. Алпатов, А. Д. Вентцель, Б. Ю. Городецкий, А. Н. Журинский, А. А. Зализняк, А. Е. Кибрик, and А. К. Поливанова. 1973. *Лингвистические задачи*. Просвещение, Москва. [V. M. Alpatov, A. D. Venttsel', B. YU. Gorodetskiy, A. N. Zhurinskiy, A. A. Zaliznyak, A. Ye. Kibrik, and A. K. Polivanova. 1973. *Lingvisticheskiye zadachi*. Prosveshcheniye, Moskva.].

А. Н. Журинский. 1993. *Слово, буква, число: Обсуждение самодостаточных лингвистических задач с разбором ста образцов жанра*. Наука, Москва. [A. N. Zhurinskiy. 1993. *Slovo, bukva, chislo: Obsuzhdeniye samodostatochnykh lingvisticheskikh zadach s razborom sta obraztsov zhanra*. Nauka, Moskva.].

А. А. Зализняк. 1963. Лингвистические задачи. In Т. Н. Молошная, editor, *Исследования по структурной типологии*, pages 137–159. Изд–во АН СССР, Москва. [A. A. Zaliznyak. 1963. Lingvisticheskiye zadachi. In T. N. Moloshnaya, editor, *Issledovaniya po strukturnoy tipologii*, pages 137–159. Izd–vo AN SSSR, Moskva.].