Electronic Language Resources in Teaching Mathematical Linguistics

Ivan Derzhanski
Department for Mathematical Linguistics
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Rositsa Dekova
Department for Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
rosdek@dcl.bas.bg

Abstract

The central role of electronic language resources in education is widely recognised (cf. Brinkley et al, 1999; Bennett, 2010; Derzhanski et al., 2007, among others). The variety and ease of access of such resources predetermines their extensive use in both research and education. With regard to teaching mathematical linguistics, electronic dictionaries and annotated corpora play a particularly important part, being an essential source of information for composing linguistic problems and presenting linguistic knowledge.

This paper discusses the need for electronic resources, especially for less studied or low-resource languages, their creation and various uses in teaching linguistics to secondary school students, with examples mostly drawn from our practical work.

1. Introduction

The mid-1960s saw the birth of the idea of presenting contemporary linguistics to secondary school students through a variety of entertaining extracurricular activities. The most prominent of those activities is the Linguistics Olympiad – contest in solving self-sufficient linguistic problems. Such problems present interesting linguistic phenomena in an enigmatic form and invite their discovery. The phenomena are presumed to be unfamiliar to the solver and may be facts of one or several languages or of language in general, or they may be ideas or concepts of linguistic science. Self-sufficiency means that a linguistic problem must be solvable using only logical thought and the information it contains, possibly supplemented by general knowledge and such concepts of linguistics, mathematics, etc., that are part of the regular school curriculum.

The First Linguistics Olympiad for secondary school students was held in 1965 in Moscow, which was the only venue of such events for 17 years. Then the linguistics competitions were launched in Bulgaria, mostly through the efforts of mathematicians, as accompanying events to contests in mathematics (Derzhanski, 2007). For these and other organisational reasons, and also because in the early years most problems that were composed in Bulgaria were on topics from mathematical or computational linguistics, linguistics as a subject of extracurricular activities for secondary school students is called 'mathematical linguistics' in Bulgaria to this day, even though the focus of the contests has shifted away from mathematical linguistics as a field of research and towards descriptive linguistics and typology. (This imprecision is tolerable, especially since, whatever topics the problems feature, the main asset for their solving is analytical thinking, which is generally associated with mathematics.)

Following similar efforts in the Netherlands and USA, in 2003 the International Olympiad in Linguistics (IOL) was launched, and has grown from 33 contestants from participating 6 countries at the first instalment to 152 contestants from 28 countries at the 12th (in 2014) and stimulated the setting up of numerous new regional and national olympiads and competitions in linguistics for secondary school students.

Thus all these countries introduced teaching contemporary linguistics (a field of study that tends to be absent from regular curricula) to secondary school students, on a narrower or broader scale, in the form of theory and practice of solving self-sufficient problems covering a wide variety of linguistic phenomena. When we refer to teaching linguistics (or mathematical linguistics) in schools in this paper, we have in mind mainly (though not exclusively) training in solving linguistic problems.

Naturally, '[a] steady supply of original, thoughtfully created and intriguing problems is absolutely necessary for the success of any ongoing [linguistic olympiad] programme' (Derzhanski and Payne, 2010). The efficiency of the problem composition and problem verification process is therefore critical. And it depends directly on the kind, size and quality of the resources available to authors and editors, especially dictionaries and corpora.

2. Language Resources for Creating Linguistic Problems

The variety and flexibility of the language resources for creating linguistic problems has to match the variety of problems, which is immense. There are monolingual problems, often on the solver's native language, focusing on little-known linguistic phenomena within the fields of grammar, semantics, or pragmatics; bilingual problems treating correspondences (regular but usually non-trivial ones) between two linguistic systems, which may be the solver's native tongue and an unfamiliar language, or the sound of a language and its written representation, or two cognate languages or dialects; and even multilingual problems, in which several such systems are compared. All levels of the language code can be involved—orthography, phonology, morphology, syntax, semantics, and discourse structure.

2.1. Use of Electronic Dictionaries

Electronic dictionaries (e-dictionaries), both monolingual and bilingual, are available now for many languages. With respect to their type and functionality, however, e-dictionaries vary widely — from a simple digital image of a printed dictionary to a digital dictionary which includes additional information (such as pronunciation or spelling in an alternative orthography, noun declension and verb conjugation, stemming and/or lemmatisation, links to derived words, sense-linked thesaurus, etc.), allows browsing, and features a powerful search engine. It is namely the latter type which serves best in composing linguistic problems, an activity in which advanced search using wildcards and/or regular expressions is especially useful.

A problem on morphology, for instance, typically illustrates some interesting rule of derivation or inflexion that makes the construction of a word or form depend on the phonological shape of the stem, the word class or some other category in a non-obvious way. To compose such a problem, one needs a significant amount of candidate data and test examples, and such can be found easily in a dictionary with adequate search tools. For example, a sizable class of Estonian nouns have single-vowel partitive plural endings, which correlate with the partitive singular ending and the stem-internal vowel. This phenomenon was demonstrated by a problem which was created using several resources: an electronic dictionary (an Estonian-Russian one) that allowed wildcard search for headwords but offered no grammatical information, the online tool Estonian Language Synthesiser[1] to verify whether the candidate words formed their partitive singular and plural forms in the required way, and a paper dictionary to resolve homonymy, which the Synthesiser doesn't do. A digital dictionary with the respective partitive singular and plural forms for every noun and an option to search for them would have made the task far easier.

Another reason to look for words of a certain morphological type may be to reduce morphological variety in a problem whose weight lies elsewhere, usually in syntax. For a problem which featured switch reference marking in Alabama the author needed to choose several verbs that would take the same set of subject and (if transitive) object person/number markers, so that the diversity of conjugation types, which is very large in this language, wouldn't obscure the main syntactic phenomenon. The verbs were collected by regular expression search in the text of an electronic edition of a paper dictionary (Sylestine et al., 1993), taking advantage of the fact that in the entries the headword was followed by grammatical

[1]Available at `http://www.filosoft.ee/gene_et/`.

information. In such cases, too, a more sophisticated structure of the dictionary can make the search significantly more efficient.

2.2. Use of Electronic Corpora

Besides dictionaries, a problem composer can use corpora as well, as tools for studying linguistic structure and as sources of naturally occurring examples of language use. Some problems are constructed entirely using material from a corpus. This is particularly desirable when the language is extinct (New Testament Greek, Middle Dutch, Tocharian, etc.) or the phenomenon calls for authentic material, as when composing problems on the structure of classical poetic forms or on word usage that occurs chiefly in literature, such as the sailors' manner of time-telling exemplified by the phrase *from about noon observation to about six bells* (Robert Louis Stevenson, *Treasure Island*). Or it may be the author's choice, aimed at making the problem more interesting. For example, a problem which presents a number of sentences in the working language which all contain the sole pronoun *we* and states that if the sentences were translated into (say) Tok Pisin, different pronouns would be used for reasons which the solver must discover, may be made more attractive if the sentences were taken from novels that the solver may know of (note that in this case it doesn't matter if the books exist in Tok Pisin at all). The quality of a corpus-based problem depends directly on the size, structure and search facilities of the corpus.

Most contemporary electronic corpora are annotated at various levels. Part-of-speech tagging is nearly ubiquitous; morphosyntactic annotation and lemmatisation is included with increasing frequency, and some corpora provide semantic and/or syntactic annotation. Furthermore, most electronic corpora are also equipped with a web search interface that allows searches for exact words or phrases, regular expressions, part of speech information, lemma, collocations, frequency and distribution of synonyms, syntactic and semantic features. These functionalities of annotated corpora and the diversity of possible queries play an essential part in contemporary problem making for the purposes of teaching mathematical linguistics.

The existence and the availability of national corpora for closely related languages, corpora of dialects or historical corpora is a useful asset for finding data for problems on phonology or morphology which draw on theoretical aspects from diachronic and comparative linguistics. Such is, for instance, a problem consisting of sentences in a regional dialect of South Bulgaria and their counterparts in contemporary standard Bulgarian where specific words are omitted so that solvers can discover a linguistic phenomenon which is present in the dialect but not in the standard (namely a distinction of proximity in demonstrative and relative pronouns and the definite article).

The availability of parallel and aligned corpora also greatly facilitates the finding of applicable excerpts of texts, as well as in the search for proper sample sentences of cognate words in unrelated languages. For example, a problem may focus on the change of meaning of cognates which could be reconstructed by students given suitable examples of natural language sentences; or students may be provided with a carefully selected coherent text and its translation and asked to discover grammar rules (a process which resembles a lot human-aided machine learning).

Problems may also comprise a set of words from two or more dialects (or closely related languages) focusing on a specific sound shift (e.g., Grimm's Law, Ruki sound law, palatalisation).

3. Task-driven Compilation of Electronic Resources

Both electronic dictionaries and corpora are often hard to come by, especially when working with exotic (or other low-resource) languages, but sometimes this difficulty can be circumvented. On one occasion, when creating a problem on Maori syntax, the author wished to have a corpus of Maori sentences in order to choose several syntactic constructions for inclusion into the problem. Since no such resource was available, a small working corpus was composed from examples given in an English–Maori dictionary (Ngata, 1993) and used successfully. Again, a large ready-made corpus with adequate search tools would have sped up the task.

Of course, not even the most sophisticated electronic dictionary or corpus can foresee all kinds of search that a user may need to perform, and the needs of authors of linguistic problems are among the most unforeseeable. It is unlikely, for example, that a dictionary will help to find anagrams, palindromic

headwords, or words which are cognate in the source and (related) target language. In such cases the problem composer (with a heart for programming) will want to download the dictionary and write his own programs to process it. Even a plain computer-readable word list is preferable to no resource at all.

4. Electronic Resources in Use by Teachers and Students

Being large and principled collections of naturally occurring language samples, corpora are used not only for composing and testing linguistic problems, but also for extracting examples to illustrate various linguistic phenomena in classroom teaching of mathematical linguistics.

When presented with a problem outside contest situations, students are usually left alone to solve the problem and thus to discover some underlying theoretical facts. Then the teacher's job is to deliver additional information on the newly discovered linguistic phenomenon and to supply examples for clarification. This is where electronic dictionaries and corpora play an essential part and help teachers provide the necessary linguistic data.

Electronic resources may be so used by students in their independent work as well. It is a recent policy of the Bulgarian Olympiad in Mathematical Linguistics that leading participants are advised to write a short research paper on a language phenomenon of their choice and to compose a sample linguistics problem (a good performance in this increases their chances to get on the national team for the International Linguistics Olympiad). And in this task students are strongly encouraged to use examples from corpora when providing linguistic evidence. Whilst originality is not expected at this stage, it is expected that the students can benefit from a small-scale first-hand encounter with linguistic research, including all stages of work with language resources (locating the resources themselves, finding the necessary information, formatting and citation). The higher accessibility of the Net, as compared to a traditional research library, means that electronic resources available online are especially well suited for this.

5. Conclusions and Future Work

In light of the rapid growth of the International Linguistics Olympiad (39 teams from 28 countries as of Edition 2014) and its national tributaries, the teaching society faces an increasing need of electronic language resources, especially on exotic and other low-resource languages, which allow for browsing and advanced searches. Although some small-size resources may be compiled in situ for a given task, the existence and the availability of large and searchable dictionaries and corpora is becoming an invaluable resource in teaching mathematical linguistics.

In the future it will be useful to establish a database with a list of available resources, as well as provide wider online access to resources created for specific teaching purposes.

Acknowledgements

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007 - 2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper.

References

Bennett, G.R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* University of Michigan Press/ELT, 144 p.

Brinkley, A., Dessants, B., Flamm, M., Fleming, C., Forcey, C. and Rothschild, E. (1999). *The Chicago Handbook for Teachers: A Practical Guide to the College Classroom, University of Chicago Press,* pages 143–67. `http://www.press.uchicago.edu/ucp/books/book/chicago/C/bo3633179.html`

Derzhanski, I.A. (2007). Extracurricular Activities in Linguistics for Secondary School Students. In Dimitrova, L. and Pavlov, R. (eds.), Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia, pages 125–128. https://www.academia.edu/5680539/Extracurricular_activities_in_linguistics_for_secondary_school_students

Derzhanski, I.A. and Payne, T. (2010). The Linguistics Olympiads: Academic Competitions in Linguistics for Secondary School Students. In Denham, K. and Lobeck, A. (eds.), *Linguistics at School: Language Awareness in Primary and Secondary Education*, Cambridge University Press, pages 213–226. https://www.academia.edu/5680870/The_Linguistics_Olym piads-Academic_competitions_in_linguistics_for_secondary_school _students

Derzhanski, I.A., Dimitrova, L. and Sendova, E. (2007). Electronic Lexicography and Its Applications: The Bulgarian Experience. In Широков, В.А. (відп. ред.), *Прикладна лінгвістика та лінгвістичні технології. Megaling-2006: Збірник наукових праць*, Київ: «Довіра», стр. 111–118.

Ngata, H.M. (1993). *English–Maori Dictionary*. Learning Media Ltd. Online version available: http://www.learningmedia.co.nz/ngata

Sylestine, C., Hardy, H.K., and Montler, T. (1993). *Dictionary of the Alabama Language.* Austin: University of Texas Press. Online version available: http://www.ling.unt.edu/~montler/Alabama/Dictionary/