Chapter 33
Language Report Slovak

Radovan Garabík

Abstract For Slovak, all the fundamental NLP building blocks for basic applications exist, but they are often of lesser quality and lower accuracy than those of other languages. The availability of free and open tools and data is rather low, with most of the resources proprietary. Compared to neighbouring languages of similar levels of NLP development (Czech, Polish, Hungarian), Slovak is positioned toward the lower end of this group. Slovak language support by "big players" in the LT industry is comparable to other European languages with similar size; speech recognition and synthesis work acceptably while machine translation between Slovak and English is almost good enough to be used by professionals as a source for post-editing. Spell checkers, LT-assisted mobile phone input, OCR and lemmatised fulltext search are taken for granted, although their quality is significantly lacking compared to bigger European languages.

1 The Slovak Language

Slovak is the official language in the Slovak Republic. Since May 2004 it has also been one of the administrative languages of the European Union. According to the 2021 census data, out of 5.4 million inhabitants of Slovakia, 4.7 million people have Slovak as their mother tongue.[1] Other estimates (perhaps overly optimistic) claim that Slovak is spoken by more than one million emigrants in the United States, about 300,000 people in the Czech Republic, and smaller groups in Hungary, Romania, Serbia, Croatia, Bulgaria, Poland and other countries. A fact which is not well known is that there exists another written variant of (Eastern) Slovak, using Cyrillic script. This variant is used around Ruski Krstur in Serbia by a few thousand speakers, but thanks to historical religious circumstances it is generally considered a dialect of the Rusyn language, not Slovak. As such, it is almost completely ignored in all aspects concerning Slovak linguistics.

Radovan Garabík
Slovak Academy of Sciences, Slovakia, radovan.garabik@kassiopeia.juls.savba.sk

[1] Corrected for the inhabitants with an unidentified mother tongue.

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_33

As a typical Slavic language Slovak is moderately inflected with a complex morphology and relatively flexible word order. It has three or four[2] genders, two grammatical numbers, three tenses and prominent aspectual pairs. It belongs (together with Polish, Czech, Lower and Upper Sorbian) to the West branch of Slavic languages. In the 16th to 18th centuries, Czech was used as the cultural language in Slovakia, together with several types of cultural Slovak, and the modern standard of the language dates to the second half of the 19th century.

Slovak is generally considered to be mutually intelligible with Czech, with some caveats regarding different inflection of pronouns, some lexical and terminological differences and differences in verb conjugations. Czech enjoys a unique sociolinguistic status in Slovakia; the population is widely exposed to the Czech language in media (TV, movies, internet, and literature). As a result, Czech is widely understood in Slovakia above the level of natural mutual intelligibility. Note that the opposite – exposure of Czech Republic inhabitants to the Slovak language – is only marginal. Despite this, the visible influence of Czech on Slovak is limited to some lexical items and syntactical constructions, often regarded as "incorrect".

The language is written using the Latin alphabet with additional diacritical marks, marking palatalisation of consonants, postalveolars, and phonemic length of vowels and consonants. The Slovak alphabet has the distinction of having the greatest number of characters (43, or 46 including digraphs) among European languages.

On the web, Slovak is a sharply localised language, closely interwoven with the .sk top-level domain (TLD). Distribution (as of 2021) of the most frequent top-level domains of web pages in the Slovak language from the Araneum Slovacum VI Maximum Beta web corpus (Benko 2014) shows that 76.6% of documents in Slovak are from the .sk TLD; 8.8% from the .com TLD, 3.8% from .cz, 2.9% from .eu, 2.0% from .net and the rest from other, less frequent domains.

2 Technologies and Resources for Slovak

Slovak language NLP and LT[3] lag behind that of neighbouring languages of similar status (i. e., Czech, Polish and Hungarian). Predominantly developed in academic environments (Šimková et al. 2012), Slovak language technologies used to be mostly limited to lemmatisation and morphosyntactic analysis, with some limited industry interest in other tools (e. g., NER). The situation has somewhat changed in recent years, with industry more interested in deep learning models. Nevertheless, the availability of huge language corpora and lexical resources available for Slovak is comparable to similar languages (Aldabe et al. 2022).

The main institution tasked with compiling and curating big, representative corpora is the Slovak National Corpus (SNK)[4] department of the Ľ. Štúr Institute of

[2] Masculine is sometimes analysed as two genders; masculine animate and masculine inanimate.

[3] See, for example, https://github.com/essential-data/nlp-sk-interesting-links

[4] https://korpus.sk

Linguistics, Slovak Academy of Sciences. SNK was also active in developing basic digital language resources of the contemporary language, but also parallel corpora, spoken, dialect and historical corpora and lexicographical databases (Garabík 2010) and in digitalisation of linguistic research in Slovakia.

Corpora compiled at SNK have formed an indispensable part of linguistic research in Slovakia for a number of years, together with the ARANEA family of huge web corpora for more than 20 languages (Benko 2014).[5] Currently, the main Slovak language corpus, prim-10.0, contains about 1.7 billion words.[6] The web corpus Araneum Slovacum VI Beta contains about 4.4 billion words. In NLP and LT industry, companies usually use in-house collected web corpora.

Official Slovak translations of various EU texts (such as Acquis communautaire, EU parliament proceedings, Official Journal of the EU etc.) make up the bulk of available, unrestricted by copyright, parallel corpora suitable for MT-related tasks.

All building blocks of basic NLP processing for Slovak are covered: lemmatisation (since Slovak is a moderately inflected language, lemmatisation is often indispensable for any subsequent language processing), and morphological analysis, including POS tagging and syntactic parsing. Spell checkers, LT-assisted mobile phone input, OCR, and lemmatised fulltext search are hidden parts of the technological background that is already taken for granted, although their quality and accuracy are lacking compared to bigger European languages. In recent years, deep learning language models appeared on the Slovak NLP scene, often adopted from comparable work for other languages (Pikuliak et al. 2021).

Recently, chatbots have noticeably penetrated many areas of human-computer interaction, as the first line of contact in customer support, and although primarily used in English-speaking countries, they are now used in other countries as well, including Slovakia, where chatbots (in written communication mostly) are used by many companies. However, since poorer accuracy of Slovak analysis leads to mixed results and the chatbots are deployed at least partly for public relations reasons, quite often these are just menu-driven FAQs (or an expert system in disguise) camouflaged by an animated head or similar graphical element, without deeper NLP processing.

3 Recommendations and Next Steps

In Slovakia, academic research and industry dealing with NLP and LT function rather separately. The academic sphere often reacts rather slowly to real demands, and instead often explores tasks with little immediate business application; the industry is mostly interested in specific tools and generally does not do NLP-related research, although there are a few companies which are active in applied NLP research.

Since many resources are not reusable due to copyright issues, clarification (i. e., opening) of the licensing of many existing datasets would be helpful for further NLP

[5] http://aranea.juls.savba.sk/aranea_about/

[6] https://korpus.sk/prim-10-0/

development. Many resources remain at the "proof of concept" stage and dedicated effort is needed to bring them up to proper levels of usability. This is also connected with the issue of sustainability of existing resources, many of which were developed as a result of specific research grants, and once the financing stopped, the resources were basically abandoned and no new development is taking place.

The Action Plan for the digital transformation of Slovakia for 2019-2022 (AP 2019) describes a centralised coordinated approach and cooperation between academic and commercial sectors in NLP. It is written only in general terms, without specific steps to be taken; the lack of computational linguists in Slovakia is not addressed (e. g., by promoting university education). The change of government after parliamentary elections in February 2020 and the COVID-19 pandemic have led to the NLP section of the Action Plan not having been acted upon at all.

References

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-strategic-documents-v3.pdf.

AP (2019). *Action plan for the digital transformation of Slovakia for 2019 – 2022*. https://www.mirri.gov.sk/wp-content/uploads/2019/10/AP-DT-English-Version-FINAL.pdf.

Benko, Vladimír (2014). "Aranea: Yet another family of (comparable) web corpora". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 247–256.

Garabík, Radovan (2010). "Slovak National Corpus tools and resources". In: *Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies*. Institute of Informatics, Slovak Academy of Sciences, pp. 2–7.

Pikuliak, Matúš, Marián Šimko, and Mária Bieliková (2021). "Cross-lingual learning for text processing: A survey". In: *Expert Systems with Applications* 165, p. 113765. DOI: 10.1016/j.eswa.2020.113765.

Šimková, Mária, Radovan Garabík, Katarína Gajdošová, Michal Laclavík, Slavomír Ondrejovič, Jozef Juhár, Ján Genči, Karol Furdík, Helena Ivoríková, and Jozef Ivanecký (2012). *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/slovak.