Chapter 34
Language Report Slovenian

Simon Krek

Abstract Around 2.5 million people around the world speak or understand Slovene, with the vast majority of them living in the Republic of Slovenia where it is the official language. The constitution grants the right to use their mother tongue to Italian and Hungarian minorities in certain municipalities. In terms of Language Technology, the Slovene CLARIN.SI consortium plays the key role in the community; all major Slovene institutions involved in the development of LT resources, tools and services are members of the consortium. In contrast, the number of private companies in Slovenia specialising in LT for Slovene remains low, and most of the LT products come either from the (Slovene) academic sphere via national or EU funding, or from the big international IT companies that cover a large number of languages.

1 The Slovenian Language

Slovene is a member of the South Slavic language family and is spoken mainly in Slovenia and the neighbouring areas in Italy, Austria, Hungary and Croatia. In the national census of 2002, the last one that recorded the number of native speakers of different languages, 87.8% of the population – of a total of just under 2 million at the time – declared Slovene to be their mother tongue, with another 3.3% claiming that they use Slovene as the language of their everyday communication at home, which amounts to 91.1% of the population using Slovene as their first language. This number puts Slovenia in the group of EU states with the most homogeneous linguistic situation. Among other linguistic groups, native speakers of languages of the former Yugoslavia were the largest in 2002, with 3.3% of them using a combination of Slovene and their mother tongue for everyday communication, and another 1% using only their mother tongue: Bosnian, Croatian, Serbian or Montenegrin. Other smaller communities included speakers of Albanian, Macedonian and Romani.

Slovene is the official language in the Republic of Slovenia. The constitution grants the right to use their mother tongue to the two minorities declaring that "in

Simon Krek
Jožef Stefan Institute, Slovenia, simon.krek@ijs.si

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_34

those municipalities where Italian or Hungarian national communities reside," Italian or Hungarian are also official languages. In 2002, it was recorded that Hungarian is the mother tongue of 0.4% of the population, and Italian of 0.2%.

According to legislation in Slovenia, all education and teaching provided as part of the current state curriculum, from preschool through to university level, must be in Slovene. In preschool, primary and secondary education, Italian is used in the schools of the Italian minority community, while Hungarian and Slovene are used in bilingual schools where the Hungarian minority is found. Special arrangements exist for children whose mother tongue is not Slovene, for the education of Roma children, children of foreign citizens and children of people without citizenship.

2 Technologies and Resources for Slovenian

A useful place to discover Slovene corpora are the CLARIN.SI NoSketch Engine[1] and KonText[2] concordancers.[3] At the time of writing, there are 76 corpora of varying sizes containing Slovene data in the repository, and 59 corpora in the concordancers. Most of them are available for download under open licences. The more important families of corpora cover general written standard language (Gigafida), Slovene Web and social media (slWaC, Janes), academic discourse (KAS), parliamentary transcriptions (siParl, ParlaMint), Slovene Wikipedia (CLASSLAWiki-sl), historical texts (IMP), literature (MAKS, ELTeC-slv), specialised domains (KoRP, DSI, Konji, etc.), and school essays (Šolar, SBSJ). There are also various manually annotated training and evaluation corpora available (ssj500k, etc.).

The GOS (GOvorjena Slovenščina, Spoken Slovene) family of corpora contains transcriptions of spoken Slovene. The original GOS includes about 120 hours of transcripts from various situations: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc.

In terms of parallel data, Slovene has benefited from its status as one of the official EU languages since 2004 and is included in the standard multilingual parallel data sets produced either by EU institutions (JRC-Acquis, DGT-Acquis, DCEP, DGT-TM, EAC-TM, ECDC-TM, JRC-Names) or by EU-funded or other projects (INTERA, WIT3, ParaCrawl, CommonCrawl, OpenSubtitles etc.), which are available either from OPUS or from repositories such as ELG. Two TM corpora produced by the Secretariat-General of the Slovene government were made available in the context of the ELRC project and are uploaded in the ELRC-SHARE repository.

There are 82 lexical/conceptual resources with Slovene data in the CLARIN.SI repository available under open access licences. Those that deserve special mention due to their size or importance are: Sloleks – morphological lexicon contain-

[1] https://clarin.si/noske/

[2] https://clarin.si/kontext/corpora/corplist

[3] https://clarin.si/info/about/

ing around 100,000 most frequent Slovene lemmas, their inflected or derivative word forms (2.7M) and the corresponding grammatical description; sloWNet is the Slovene WordNet developed in the expand approach: it contains the complete Princeton WordNet 3.0 and over 70,000 Slovene literals; Dictionary of the Slovenian Normative Guide is a normative orthographic dictionary of Slovene standard language. It contains 140,266 lemmas and sublemmas in 92,617 entries; Thesaurus of Modern Slovene is an automatically created thesaurus from Slovene data available in a comprehensive English–Slovene dictionary, a monolingual dictionary, and a corpus. It contains 105,473 entries and 368,117 synonym pairs.

In terms of language models, the most recent one is the Slovene RoBERTa model. The corpora used for training the model contain 3.47 billion tokens in total. The subword vocabulary contains 32,000 tokens.[4] Multilingual models are also available, e. g., a trilingual BERT model, trained on Croatian, Slovene, and English data.[5]

The standard and most accurate text processing tool for Slovene is the CLASSLA fork of the Stanza pipeline.[6] It supports processing of both standard and non-standard Slovene at the level of tokenisation and sentence segmentation, part-of-speech tagging, lemmatisation, dependency parsing and named entity recognition.

There are some Slovene LT companies that develop speech-to-text and text-to-speech tools.[7] Slovene is also available in speech technology services offered by large enterprises such as Microsoft and Google, as well as by other companies specialising in speech technology.[8] These solutions have also found their way into some specialised devices covering many languages.[9] At the University of Ljubljana, a system has been developed for automatically translating lectures from Slovene to other languages in real time, in the context of the Online Notes project.[10]

Machine translation services for Slovene are available through more or less the same stakeholders: some Slovene LT companies,[11] the large enterprises such as Microsoft and Google, and some other international companies specialising in machine translation technology or general translation services.[12] As an official EU language, Slovene is included in the eTranslation service offered by the European Commission.

The biggest investment in LT for Slovene is the Development of Slovene in Digital Environment project financed by the Slovene Ministry of Culture between 2020-2023.[13] The project will significantly upgrade existing LT resources, tools and services, or produce many of those that do not exist yet. The results of the project are

[4] http://hdl.handle.net/11356/1397

[5] http://hdl.handle.net/11356/1330

[6] https://github.com/clarinsi/classla, https://pypi.org/project/classla/

[7] Amebis, Alpineon: eBralec, https://ebralec.si; Vitasis: Truebar, https://vitasis.si

[8] NEWTON Technologies, https://www.newtontech.net; Sonix: https://sonix.ai

[9] Pocketalk: https://europe.pocketalk.com/languages-countries/

[10] https://www.cjvt.si/en/infrastructure-support/tolmac/

[11] Vitasis: Truebar, https://vitasis.si; Aikwit, https://aikwit.com; Taia, https://taia.io

[12] DeepL Translate, https://www.deepl.com; Pangeanic, https://pangeanic.com/languages/slovenian-translation-services/, etc.

[13] Razvoj slovenščine v digitalnem okolju (RSDO): https://www.slovenscina.eu

expected to be published on the CLARIN.SI and GitHub repositories in November 2022 and February 2023.

3 Recommendations and Next Steps

In general, one can conclude that 1. the support for Slovene is comparable with other languages with a similar status (Krek 2022, 2012), 2. there is a general awareness in governmental bodies that LT for Slovene should be supported in the future, 3. the LT community is growing, also through new educational initiatives such as the MA study of Digital Linguistics (Faculty of Arts, University of Ljubljana), and 4. there is infrastructural support, mainly through the CLARIN.SI infrastructure at the Jožef Stefan Institute, which also covers all other stakeholders through the CLARIN.SI consortium. However, more efforts are needed in the future to bring the existing support closer to those available for other (official EU) languages.

References

Krek, Simon (2012). *Slovenski jezik v digitalni dobi – The Slovene 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/slovene.

Krek, Simon (2022). *Deliverable D1.31 Report on the Slovenian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-slovenian.pdf.