EUROPEAN
LANGUAGE
EQUALITY

D1.30

Report on the Slovak Language

Author	Radovan Garabík
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.30
Deliverable title	Report on the Slovak Language
Type	Report
Number of pages	20
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Author	Radovan Garabík
Reviewers	Jaroslava Hlavacova, Guðrún Gísladóttir
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2022 ELE Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CEF AT	Connecting Europe Facility, Automated Translation
CL	Computational Linguistics
DLE	Digital Language Equality
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
GPU	Graphics Processing Unit
HPC	High-Performance Computing
LR	Language Resources/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLG	Natural Language Generation
NLP	Natural Language Processing
OCR	Optical Character Recognition
SMS	Short Message Service
SR	Speaker Recognition
TLD	Top Level Domain

Abstract

Processing natural human language by computers is a complex and non-trivial task that used to be an elusive goal in research and industry for decades, with only partial and imperfect solutions and slow progress. It has been established as a specialised scientific field known as *Computational Linguistics*, *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). Often considered a subfield of Artificial Intelligence, modern language technologies benefit from recent advances in the field, especially with deep learning combined with the availability of large language data and accessible computing power in the form of GPUs. Language technology is 'behind the scenes' in many areas of our daily life, starting with predictive input on the virtual keyboards on mobile phones, with spell checkers, using internet search machines but also elsewhere in dealing with either huge language data or human-computer interaction. However, there is a noticeable gap in the level of support for different languages and in the availability of resources for training NLP tools. Since most of the research and development is performed for/within English, and there is a lot of freely reusable English language content available, English has the best support and most of the state-of-the-art methods and tools are developed for and in English. Then follows a group of 'big European' languages – German, French, and Spanish with good support; most of other (national) European languages have fragmentary support with reasonably developed basic NLP tools and enough available data; 'smaller' languages, such as Irish or Maltese struggle to maintain even weak support for NLP. Then there is a group of languages, usually minority or endangered ones with none (or almost none) resources, such as Rusyn or Romani.

Although Slovak belongs to the group with fragmentary support, its position is toward the lower end of the group, relative to other comparable languages, e. g., Czech, Polish or Hungarian. For Slovak, all the fundamental NLP building blocks necessary for basic applications are present, but they are often of lesser quality and achieving lower accuracy, sometimes barely advancing beyond the proof of concept stage, and there is much less choice between different tools performing similar tasks – often there is only one implementation available. The availability, especially of free and open tools and data is also rather low, with most of the resources proprietary.

Slovak language support by "big players" in the LT industry is comparable to other European languages with similar size – speech recognition and synthesis works acceptably, machine translation between Slovak and English (translation from/to other languages is of lower quality) is almost good enough to be used by professional translators as a source for post-editing. Spelling checkers, LT assisted mobile phone input, OCR, lemmatised fulltext search are already taken for granted, although their quality and accuracy significantly lacks compared to bigger European languages.

The status of Slovak as a language with less developed NLP resources is especially striking when compared with Czech, which enjoys excellent research (at the top European level) and consequently the best LT support among Central European Languages, whereas Slovak ranks the worst in this group.

Rozšírený abstrakt

Spracovanie prirodzeného jazyka je zložitá a netriviálna úloha, ktorá sa celé desaťročia vnímala ako nedosiahnuteľný cieľ vedeckého výskumu s minimálnymi možnosťami zavedenia do praxe, ako disciplína s čiastočnými a nedokonalými riešeniami a pomalými pokrokmi. Danej problematike sa venujú špecializované vedné oblasti, a to *počítačová lingvistika* a *počítačové spracovanie prirodzeného jazyka* (NLP), pod súhrnným názvom známe aj ako *jazykové technológie* (LT). Moderné jazykové technológie, ktoré sa často považujú za podoblasť umelej

inteligencie (AI), profitujú z novodobého prudkého rozvoja technológií AI, najmä z techniky strojového učenia (deep learning) v kombinácii s dostupnosťou veľkých jazykových korpusov a výpočtovým výkonom vo forme GPU. Jazykové technológie sa využívajú v mnohých oblastiach každodenného života, napríklad pri prediktívnom písaní na virtuálnych klávesniciach mobilných telefónov, kontrole pravopisu, používaní internetových vyhľadávačov, ale aj pri práci s rozsiahlymi textovými dátami a všeobecne v interakcii medzi človekom a počítačom.

Úroveň podpory jazykových technológií a dostupnosť jazykových zdrojov použiteľných na trénovanie nástrojov na spracovanie prirodzeného jazyka je pre rôzne jazyky značne rozdielna. Výskum a vývoj jazykových technológií prebieha predovšetkým so zameraním sa na angličtinu, navyše je v tomto jazyku k dispozícii množstvo voľne použiteľných zdrojov, angličtina má preto prirodzene najlepšiu podporu a je pre ňu vyvinutá väčšina najmodernejších metód a nástrojov. Nasleduje skupina „veľkých európskych" jazykov s dobrou podporou, ktorú tvorí nemčina, francúzština a španielčina. Väčšina ostatných (národných) európskych jazykov má čiastočnú podporu s primerane vyvinutými základnými nástrojmi NLP a dostatkom dostupných údajov. Ale „menšie" jazyky, ako napríklad írsky alebo maltský, majú problém udržať krok hoci aj so slabou podporou NLP. Napokon existuje skupina prevažne menšinových alebo ohrozených jazykov, ktoré nemajú žiadne, resp. takmer žiadne zdroje a nástroje NLP, patrí tu napríklad rusínčina alebo rómčina.

Hoci slovenčina patrí do skupiny s čiastočnou podporou jazykových technológií, v konfrontácii s inými porovnateľnými jazykmi, akými sú čeština, poľština či maďarčina, zaostáva a jej pozícia je medzi stredoeurópskymi jazykmi až na poslednom mieste. Pre slovenčinu sú k dispozícii všetky stavebné bloky NLP potrebné na tvorbu základných aplikácií, ale tieto sú často menej kvalitné a dosahujú menšiu presnosť v porovnaní s inými jazykmi, niekedy sotva dosahujú úroveň funkčných prototypov; pre slovenčinu je tiež zúžený výber nástrojov vykonávajúcich podobné úlohy, často je k dispozícii iba jedna implementácia. Dostupnosť najmä bezplatných a otvorených nástrojov a dát je tiež pomerne nízka, pretože väčšina zdrojov je uzavretá, bez možnosti verejného prístupu.

Podpora slovenčiny zo strany „veľkých hráčov" v oblasti priemyselného využitia LT je približne rovnaká ako pre ostatné porovnateľné európske jazyky: rozpoznávanie a syntéza reči fungujú prijateľne spoľahlivo, strojový preklad medzi slovenčinou a angličtinou (preklady z/do iných jazykov sú menej spoľahlivé) je už použiteľný aj profesionálnymi prekladateľmi ako východiskový zdroj pre následné manuálne úpravy. Kontrola pravopisu, asistovaný vstup textu z mobilného telefónu, optické rozpoznávanie znakov, lemmatizované fulltextové vyhľadávanie sú už samozrejmosťou, aj keď ich kvalita a presnosť výrazne zaostáva za väčšími európskymi jazykmi.

Pozícia slovenčiny ako jazyka s menej rozvinutými zdrojmi NLP je najvýraznejšia hlavne v porovnaní s češtinou. Výskum českého jazyka je na špičkovej európskej úrovni a má najlepšiu podporu spomedzi stredoeurópskych jazykov, pričom slovenčina má v rámci tejto skupiny jazykov, ako sme uviedli vyššie, najslabšiu podporu LT.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed in the frame of the European Language Equality (ELE) project. With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

2 The Slovak Language in the Digital Age

2.1 General Facts

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 2011 census data,[2] out of 5.4 million inhabitants of Slovakia, 4.6 million people have Slovak as their mother tongue, 4.7 million use Slovak as their primary language in public and 4.5 million as their primary language in private.[3]

Other estimates (perhaps overly optimistic) claim that Slovak is spoken by more than 1 million emigrants in the United States, and approx. 300,000 people in the Czech Republic. Smaller language groups of Slovaks are situated 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 (Serbia) by a few thousand speakers, but thanks to historical religious circumstances it is generally considered a dialect of the Ruthenian language, not Slovak. As such, the language development and use is disconnected from the language of the Slovak Republic and is almost completely ignored in all aspects concerning Slovak linguistics.

Slovak belongs (together with Polish, Czech, Lower and Upper Sorbian) to the West branch of Slavic languages. The Proto-Slavic basis of Slovak was formed in the area between the Carpathians, the Danube, and the Upper Moravia. The Slovak language went through fast development in the 10th to 12th centuries and stabilised in the 13th to 15th centuries. In the 16th to 18th centuries, Czech was used as the cultural language in Slovakia, together with several types of cultural Slovak. By the end of the 18th century, attempts at the formation of literary Slovak had started. At the end of the 18th century, Anton Bernolák based his codification on cultural West Slovak, but failed to get wider recognition. Ľudovít Štúr used Central Slovak as the basis and his idea took hold very soon, and with certain modifications (by Martin Hattala and Michal Miloslav Hodža) lasts up to these days, with the last significant orthography reform[4] in 1953. The literary (standardised) Slovak is thus a relative latecomer among European languages.

Slovak is generally considered to be mutually intelligible with Czech, with some caveats regarding different inflection of pronouns, lexical differences (most prominent in culinary terms, botanical and zoological taxonomy) and differences in verb conjugations.[5] Czech enjoys unique sociolinguistic status in Slovakia – the population is widely exposed to the Czech

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

[2] The data from the 2021 census are not available at the time of writing.

[3] These numbers are corrected by an estimated ratio of "undetermined" language users – the census has been marked by people selecting various other values as a way of protesting privacy issues surrounding the process.

[4] Pre-1953 orthography is sufficiently different to significantly impair modern Natural Language Processing (NLP) tools if used on older texts

[5] Especially compared to Colloquial Czech; literary Czech is closer to Slovak in this regard.

language in media (Czech TV, movies, internet; literature in Czech is widely read, though in decline), and as a result of this exposure Czech is widely understood in Slovakia above the level of natural mutual intelligibility (the opposite – exposure of Czech Republic inhabitants to the Slovak language – is only marginal). Despite this, visible influence of Czech on Slovak is limited to some lexical items and syntactical constructions formally regarded as "incorrect".

Slovak as a typical Slavic language is a moderately inflected language with a complex morphology and relatively flexible word order. It has three or four[6] genders, two grammatical numbers, three tenses and prominent aspectual pairs.

The language is written using the Latin alphabet with additional diacritical marks. The palatalisation of consonants is marked by a háček (*ď*, *ť*, *ň*, *ľ*; háček is also used for unrelated graphemes *ž*, *š*, *č*, (*dž*), representing postalveolar consonants) and the length of vowels and consonants by an acute accent (*á*, *é*, *í*, *ó*, *ú*, *ý*, *ĺ*, *ŕ*). Diaeresis/umlaut is used in the letter *ä* and circumflex in *ô*. Letters *ö* and *ü*, while not being formally part of the alphabet, are also marginally used in the standard orthography. The Slovak alphabet has the distinction of having the greatest number of characters (43; or 46 including digraphs) among European languages.

A common way (though declining) of writing Slovak in an environment without proper support of the Slovak alphabet (e. g., on pre-Unicode environment, on the internet, in SMS messages, previously in telegram messages etc.) is to drop the diacritics (the letter *ä* is sometimes changed into *e*).

2.2 Slovak in the Digital Sphere

On the web, Slovak is a sharply localised language – closely interwoven with the `.sk` top level domain (before the advent of generic TLDs). Distribution of most frequent top level domains of webpages in the Slovak language is show in the Table 1, data is from the *Araneum Slovacum VI Maximum Beta* web corpus (Benko, 2014) as of 2021.

TLD	% of documents
.sk	76.6
.com	8.8
.cz	3.8
.eu	2.9
.net	2.0
.org	1.8
.info	1.3

Table 1: Distribution of top level domains of Slovak language webpages

There were 4.64 million internet users in Slovakia in January 2021, which is an increase of 2.4% since 2020, Internet penetration in Slovakia was at 85.0% in January 2021. The number of social media users was 73.8% of the total population in January 2021, which is an increase of 11% since 2020 (DataReportal, 2021).

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

3 What is Language Technology?

Natural language[7] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc., which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e., systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.

[7] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e., the generation of speech given a piece of text, Automatic Speech Recognition, i. e., the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e., the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e., the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[8]

4 Language Technology for Slovak

Overall, Slovak language NLP[9] and Language technologies are lagging behind neighbouring languages of similar status (e. g., Czech, Polish and Hungarian). Being predominantly developed in academic environment (Šimková et al., 2012), in the past the language technologies

[8] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[9] An updated list of interesting NLP resources for Slovak can be found at https://github.com/essential-data/nlp-sk-interesting-links

were mostly limited to lemmatisation and morphosyntactic analysis, with some limited industry development mostly leading to the development and use of Named Entity Recognition as an important component of industrial NLP. The situation somewhat changed in recent years, with the industry more interested in deep learning models (trained on web corpora). On the other hand, huge language corpora and lexical resources availability is comparable to similar languages.

4.1 Language Data and Tools

For many years, the main institution collecting and curating language data and tools for processing Slovak text has been the Slovak National Corpus[10] department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. The project was primarily aimed at building basic electronic language resources of the contemporary Slovak language, especially building of a huge representative corpus, but also parallel corpora, spoken, dialect and historical corpora and lexicographical databases (Garabík, 2010). The project also catered to digitalisation of linguistic research in Slovakia. The corpora in the Slovak National Corpus collection form a valuable data source for all levels of language processing; however, most of them share the common problem of corpus linguistics – the inability to redistribute the data because of copyright restrictions. At the time of writing, the main Slovak language corpus *prim-9.0* contains about 1.6 billion words;[11] in addition, a web corpus *Araneum Slovacum VI Beta*[12] contains about 4.4 billion words (similar sizes are reported from other teams building Slovak web corpora).

Parallel corpora compiled by the Ľ. Štúr Institute of Linguistics are aimed at languages important in the Slovak environment. The biggest and most developed corpora are therefore Slovak↔English and Slovak↔Czech, with Slovak↔Russian, Slovak↔German following, and several other small parallel corpora for other languages (see Vasilišinová and Garabík (2009); Dimitrova and Garabík (2011); Garabík (2015)).

Since Slovak was recognised as an official EU language in 2004, official translations of various EU texts (such as Acquis communautaire, EU parliament proceedings, Official Journal of the EU etc.) make the bulk of available, unrestricted by copyright parallel corpora or translation memories. Although primarily translations from English or French, due to their nature, these corpora contain pairs of almost all official EU languages and their size is sufficient for many NLP and MT related tasks, although the language of the corpora is rather monothematic and limited in covered domains, vocabulary and style. The uneven coverage of domains in freely available (for various definitions of the word) corpora is being at least partly addressed within the *CURLICAT – Curated Multilingual Language Resources for CEF AT* project.[13]

All the building blocks of basic NLP processing for Slovak are already covered – lemmatisation, full morphological analysis, including part of speech tagging (Garabík and Bobeková, 2021), syntactic parsing (Straka and Straková, 2017). In recent years, deep learning language models started to appear on the Slovak NLP scene, often adopted from comparable work for other languages (Pikuliak et al., 2021a).

In general, the percentage of free and open resources is still rather low – huge corpora were (and are) copyright encumbered, and existing tools were often closed, albeit sometimes available commercially. Nevertheless, the (n-gram) language models trained on the Slovak National Corpus are publicly available,[14] as well as other smaller tools, resources

[10] https://korpus.juls.savba.sk

[11] https://korpus.sk/prim(2d)9(2e)0.html

[12] http://aranea.juls.savba.sk/aranea/run.cgi/corp_info?corpname=AranSlov_a

[13] https://curlicat.eu

[14] https://korpus.sk/prim(2d)7(2e)0(2f)models.html

and corpora without copyright restrictions. Recently, Slovak BERT model trained on a web corpus has been released without any restrictions by the Kempelen Institute of Intelligent Technologies and Gerulata Technologies (Pikuliak et al., 2021b).

Slovak is a moderately inflected language, with most of the inflections realised via suffixes, however, these suffixes often interact with the root morpheme of the word in unpredictable ways, there is a sizeable amount of homonymy of the suffixes and a non-negligible amount of additional changes in the root of the word. This precludes creation of simple rule based stemming algorithms, and altogether makes stemming somewhat impractical to implement. Therefore various full text search engines (as a basic prerequisite for information extraction) usually make use of full lemmatisation lists, without disambiguation (thus increasing recall at the expense of precision). Efficient stemming can still be trained on the lemmatisation data (see the Lucene/SOLR implementation below).

There is support for Slovak in several most popular search engines: there is a Slovak stemmer implementation for Lucene/SOLR,[15] lemmatizer for Lucene,[16] Slovak support for Elastic search.[17] Such lemmatisation lists are either based on Slovak *ispell* or *aspell* data,[18] or use morphological database developed at the JÚĽŠ SAV.

In recent years, chatbots noticeably penetrated many areas of human-computer interaction. They are especially prominent in the business sphere in customer support, as the "first line" of contact, and although primarily used in English speaking countries, the chatbots are now widespread in other countries as well. Slovakia is no exception, although the chatbot hype declined somewhat, chatbots (in written communication mostly) are commonly used by many companies in commercial environment. However, because poorer accuracy of Slovak analysis leads to mixed results and the chatbots are deployed at least partly because of public relations reasons, quite often the "chatbots" are just menu driven FAQs (or an expert system in disguise) camouflaged by an animated head or similar graphical element, without any deeper NLP processing.

Although document summarisation has been recently getting more attention, Slovak language support and research in this area has been significantly undervalued. However, recently there was some effort in producing Slovak language dataset for summarisation (Suppa and Adamec, 2020) (although the availability of the data is unknown). There are also ongoing works regarding language generation, (e. g., Blšták and Rozinajová (2021); Vasko et al. (2020)), but the effort is scarce and results only preliminary.

4.2 Projects, Initiatives, Stakeholders

Slovak National Corpus is a catch-all name of a set of four consecutive national projects (2002-present), carried out by the eponymous department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. Originally aimed at building a comprehensive, representative big (national) corpus of written Slovak and to bring contemporary computer technologies into linguistic research in Slovakia, the project(s) later continued with creating tools and databases for basic NLP tasks, compiling bilingual corpora, corpus of spoken language and specialised corpora (corpus of dialects, historical corpora, etc.).

Speech recognition and synthesis for Slovak were pioneered by the Speech analysis and synthesis department of the Institute of Informatics, Slovak Academy of Sciences.[19] Speech recognition was first deployed on a massive scale within the project of automatic dictation transcription system for the Ministry of Justice of the Slovak Republic (Rusko et al., 2016).

[15] https://github.com/essential-data/stemmer-sk

[16] https://github.com/essential-data/lucene-fst-lemmatizer

[17] https://github.com/essential-data/elasticsearch-sk

[18] http://spell.linux.sk

[19] http://www.ui.sav.sk/pp/speech/

There are also efforts to extend speech recognition to regional variants and dialects of Slovak (Darjaa et al., 2018), an area often neglected in NLP processing.

There is an *Action Plan for the digital transformation of Slovakia for 2019 – 2022*[20] that contains a specific section on NLP. There is a section describing the plans to create a unified, centralised and coordinated approach and provide for cooperation between academic and commercial sectors. However, it is written in very general terms, without any specific steps to be taken. The financing source is given as "EU Funds" without any further specification – and that likely means national funding was not planned. The new government/administration after parliamentary elections in February 2020 inherited this agenda (that they did not prepare), and combined with the COVID-19 pandemic resulted in the NLP section of the Action Plan not being acted upon at all. The deadline described in the plan – the end of 2020 – has not been met. The Action Plan does not address the lack (compared to other countries/languages) of computational linguists in Slovakia at all, e. g., by suggesting to create university education.

The second relevant plan is the *Strategy of the Digital Transformation of Slovakia 2030*.[21] The general aim is to create a functional digital economy in Slovakia. The main focus remains at e-health services, cybersecurity and data protection, with further improvement in e-Government development. The agenda plans and relies on EU-wide integration and connectivity.

5 Cross-Language Comparison

The LT field[22] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[23] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)

[20] https://www.mirri.gov.sk/wp-content/uploads/2019/10/AP-DT-English-Version-FINAL.pdf

[21] https://www.mirri.gov.sk/wp-content/uploads/2019/11/Brochure-SMALL.pdf

[22] This section has been provided by the editors.

[23] Tools tagged as "language independent" without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

  - Human-computer interaction (e. g., tools for conversational systems)

- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[24]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[25] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not

[24] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

[25] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[26]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 2 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[27] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning,

[26] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

[27] In addition to the languages listed in Table 2, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

			Tools and Services							Language Resources					
			Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	**Overall**
EU official languages		Bulgarian	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Croatian	▒	░	░	▒	▒	▒	░	▒	░	▒	░	▒	▒
		Czech	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	░	▒	▒
		Danish	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	▒	▒
		Dutch	▒	▓	▓	▒	▓	▓	▓	▒	░	▒	▒	▒	▒
		English	█	█	█	█	█	█	█	█	▓	█	▓	█	█
		Estonian	▒	▒	▓	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Finnish	▒	▒	▓	▒	▒	▓	▓	▒	▒	▒	▒	▒	▒
		French	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▒	▓	▓
		German	▓	█	█	█	█	█	█	▓	▓	▓	▒	▓	▓
		Greek	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	▒	▒	▒
		Hungarian	▒	▒	▒	▒	▒	▒	▓	▒	░	▒	▒	▒	▒
		Irish	░	░	░	░	░	▒	░	▒	░	▒	░	░	░
		Italian	▒	▓	▓	▓	▒	▓	▓	▒	▒	▒	░	▒	▒
		Latvian	▒	▒	▓	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Lithuanian	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Maltese	░	░	░	░	░	▒	░	░	░	▒	░	░	░
		Polish	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	░	▒	▒
		Portuguese	▓	▓	▓	▒	▒	▓	▓	▒	▒	▒	░	▒	▒
		Romanian	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	▒	▒
		Slovak	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	░	▒
		Slovenian	▒	▒	▒	▒	▒	▒	░	▒	░	▒	░	▒	▒
		Spanish	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓
		Swedish	▒	▒	▒	▒	▒	▓	▓	▒	▒	▒	░	▒	▒
(Co-)official languages	National level	Albanian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Bosnian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Icelandic	▒	▒	▒	░	░	▒	▒	░	░	▒	░	░	▒
		Luxembourgish	░	░	░	░	░	▒	░	░	░	░	░	░	░
		Macedonian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Norwegian	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	░	▒
		Serbian	░	░	▒	▒	▒	▒	░	░	░	░	░	░	░
	Regional level	Basque	▒	▒	░	▒	░	▒	▒	░	░	░	░	░	▒
		Catalan	▒	▒	▒	▒	▒	▓	▒	▒	░	░	▒	▒	▒
		Faroese	░	░	░	░	░	░	░	░	░	░	░	░	░
		Frisian (Western)	░	░	░	░	░	░	░	░	░	░	░	░	░
		Galician	▒	░	▒	▒	░	▒	▒	░	░	░	░	░	▒
		Jerriais	░	░	░	░	░	░	░	░	░	░	░	░	░
		Low German	░	░	░	░	░	░	░	░	░	░	░	░	░
		Manx	░	░	░	░	░	░	░	░	░	░	░	░	░
		Mirandese	░	░	░	░	░	░	░	░	░	░	░	░	░
		Occitan	░	░	░	░	░	░	░	░	░	░	░	░	░
		Sorbian (Upper)	░	░	░	░	░	░	░	░	░	░	░	░	░
		Welsh	░	▒	░	░	░	▒	▒	░	░	░	▒	░	▒
		All other languages	░	░	░	░	░	░	░	░	░	░	░	░	░

Table 2: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

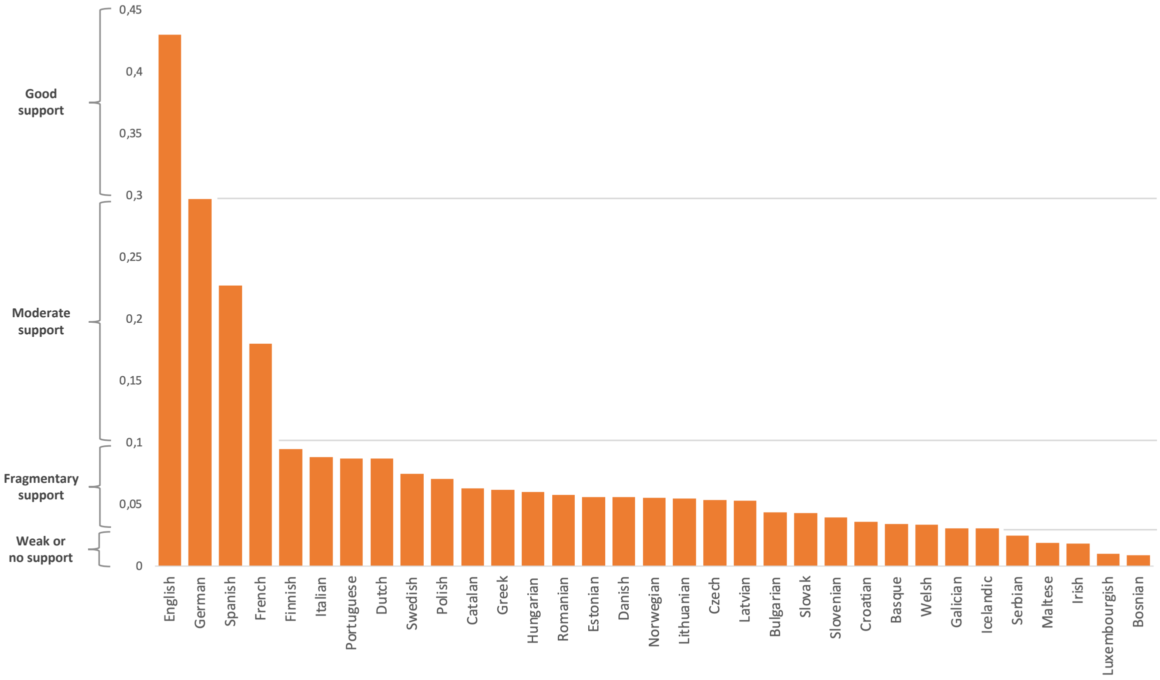

Figure 1: Overall state of technology support for selected European languages (2022)

emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

Slovak language support by “big players” in the LT industry is comparable to other European languages with similar size – speech recognition and synthesis works acceptably, machine translation between Slovak and English (translation from/to other languages is of lower quality) is on the verge of being good enough to be used by professional translators as a source for post-editing (at least for general domain). Spelling checkers, LT assisted mobile phone input, OCR, lemmatised fulltext search are parts of hidden technological background that is already taken for granted, although their quality and accuracy significantly lacks compared to bigger European languages (while English support is yet in quite another league).

Big monolingual Slovak language corpora provided by the Ľ. Štúr Institute of Linguistics are an indispensable part of linguistic research in Slovakia for a number of years, together with the ARANEA family of comparable huge web corpora for more than 20 languages[28] (Benko, 2014); in commercial settings, business oriented NLP and language technologies, companies usually use in-house collected web corpora. While the language content of web corpora is usually of lower quality, web corpora are bigger and with some care and filtering they are adequate for many uses, and can usually be built in-house with just a moderate effort, which alleviates very important copyright issues. The size of available corpora is already sufficient for most practical uses.

While not strictly NLP, lexical resources of the Dictionary Portal[29] of the Ľ. Štúr Institute of Linguistics provides several dictionaries and dictionary-like databases, including basic normative dictionaries, modern corpus-based and corpus-driven dictionaries, both scholarly and public-oriented, and an access to several corpora and language resources presented in a unified dictionary-like format. It is an invaluable resource for many kinds of linguistic research and a site for general dictionary consulting. Given the popularity of the portal, this is often the first place that people interested in language technologies visit and see distilled results of NLP in lexicography.

Compared to other similarly sized European languages, there is less variety of existing Slovak language tools and resources for a given task, and/or lower accuracy and coverage, and many existing resources are barely advanced beyond a prototype stage quality. This is especially visible when compared with LT support for Czech, which is among the top European players.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

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

Miroslav Blšták and Viera Rozinajová. Automatic question generation based on sentence structure analysis using machine learning approach. *Natural Language Engineering*, page 1–31, 2021. doi: 10.1017/S1351324921000139.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

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

[29] https://slovnik.juls.savba.sk

Sakhia Darjaa, Róbert Sabo, Marián Trnka, Milan Rusko, and Gabriela Múcsková. Automatic Recognition of Slovak Regional Dialects. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 305–308, 2018. doi: 10.1109/DISA.2018.8490639.

DataReportal. Digital 2021 Slovakia. https://datareportal.com/reports/digital-2021-slovakia, 2021. Accessed: 2021-12-08.

Ludmila Dimitrova and Radovan Garabík. Bulgarian–Slovak Parallel Corpus. In *Natural Language Processing, Multilinguality. Proceedings of the 6th International Conference SLOVKO 2011*. Tribun, Brno, 2011.

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

Radovan Garabík. Slovak-English Parallel Corpus. In *Semantyka a konfrontacja językova*, pages 117–123. Slawistyczny Ośrodek Wydawniczy Instytutu Slawistyki PAN, 2015. ISBN 978-83-64031-25-0.

Radovan Garabík and Kristína Bobeková. Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. *Slovenská reč*, 86(1):104–109, 2021.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765, 2021a. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113765. URL https://www.sciencedirect.com/science/article/pii/S0957417420305893.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. SlovakBERT: Slovak Masked Language Model, 2021b.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Milan Rusko, Jozef Juhár, Marián Trnka, Ján Staš, Sakhia Darjaa, Daniel Hládek, Róbert Sabo, Matúš Pleva, Marian Ritomský, and Stanislav Ondáš. Advances in the Slovak Judicial Domain Dictation System. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9561:55–67, 2016.

Milan Straka and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Marek Suppa and Jergus Adamec. A Summarization Dataset of Slovak News Articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6725–6730, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.830.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Dorota Vasilišinová and Radovan Garabík. Parallel French-Slovak Corpus. In *Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007*. Tribun, Brno, 2009.

Dominik Vasko, Samuel Pecár, and Marián Šimko. *Automatic Text Generation in Slovak Language*, pages 639–647. Springer International Publishing, 01 2020. ISBN 978-3-030-38918-5. doi: 10.1007/978-3-030-38919-2_53.

Mária Šimková, Radovan Garabík, Katarína Gajdošová, Michal Laclavík, Slavomír Ondrejovič, Jozef Juhár, Ján Genči, Karol Furdík, Helena Ivoríková, and Jozef Ivanecký. *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012. URL http://www.meta-net.eu/whitepapers/volumes/slovak. Georg Rehm and Hans Uszkoreit (series editors).