EUROPEAN
LANGUAGE
EQUALITY

D1.8

Report on the Czech Language

Author	Jaroslava Hlavacova
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	D1.8
Deliverable title	Report on the Czech Language
Type	Report
Number of pages	23
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	Jaroslava Hlavacova
Reviewers	Annika Grützner-Zahn, Maria Giagkou
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

ACL	Association of Computational Linguistics
AI	Artificial Intelligence
ASRU	Automatic Speech Recognition and Understanding
BDVA	Big Data Value Association
CL	Computational linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DARPA	Defense Advanced Research Projects Agency
EIA	Environmental Impact Assessment
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)
ELITR	European Live Translator
ELRA	European Language Resources Association
ERIC	European Research Infrastructure Consortium
ESIC	Europarl Simultaneous Interpreting Corpus
EU	European Union
GPU	Graphics Processing Unit
GW	Giga-word
HPC	High-Performance Computing
IARPA	Intelligence Advanced Research Projects Activity
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
ISCA	International Speech Communication Association
JSON	JavaScript Object Notation
LDC	Linguistic Data Consortium
LINDAT/CLARIAH-CZ	Czech centre for data providing certified storage and natural language processing services
ICASSP	The international Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
LDC	Linguistic Data Consortium
LT	Language Technology
META-NET	EU Network of Excellence to foster META
ML	Machine Learning

MT	Machine Translation
MW	Mega-word
NLG	Natural Language Generation
NLP	Natural Language Processing
NSF	National Science Foundation
ONR	Office of Naval Research
PDT	Prague Dependency Treebank
POS	Part of Speech

Abstract

This report belongs to the series of reports worked out for about 30 European languages. It is one of the outputs of the European Language Equality (ELE) project. The objective of the series is twofold:

- to outline the current status of the NLP for a given language;
- to draw attention to possible gaps or stagnations.

In the field of natural language processing (both text and speech processing), the Czech Republic is considered a great power in the international scientific community. In 2021 it hosted, amongst other things, the largest conference in the field of speech technology – Interspeech. In the past, the Czech Republic organised several major events in the field of language processing in general – world congress of Association for Computational Linguistics, COLING, international conferences IEEE ICASSP, ISCA Odyssey and IEEE ASRU.

This report delivers the basic data about NLP for the Czech language. After a brief introduction (Section 1) with general facts about the language (history, basic linguistic features, writing system, dialects), the report focuses on the presence of Czech in the digital sphere (Section 2). The main achievements in the field of NLP are presented in Section 4.1, together with examples of the most important datasets (corpora, treebanks, lexicons etc.) and tools (morphological analyzers, taggers, automatic translators, voice recognisers and generators, keyword extracters etc). This section also contains information about the most prominent projects, initiatives and stakeholders (Section 4.3, especially the most recent ones (the last 5 years). The list of references at the end of the report leads to detailed description of all the documents, datasets and tools mentioned in the text.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support these 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.

[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 Czech Language in the Digital Age

2.1 General Facts

Czech, one of the West Slavonic languages, has about 10 million speakers. Most of them live in the Czech Republic (also called Czechia). In other parts of the world, there are about 200 thousand speakers, mostly emigrants and their descendants. The Czech language is an official language in the Czechia, since May 2004 it is also one of the administrative languages of the EU. Czech is used during administrative, judicial and other official proceedings. The manuals and description of imported goods must contain their Czech translation.

The Czech language has several varieties, especially in its spoken form. Literary (standard) Czech is a prestige variety, which is taught in schools and strongly preferred in official texts and in mass media. However, literary Czech is not prescribed by any law. In everyday communication, most people prefer other varieties of Czech. The most widespread one is common Czech, based on the Central Bohemia dialects. In Moravia and Silesia, the dialects such as Hanak, Lach, Czecho-Moravian, are still used actively in spoken form. Common Czech and dialects differ from the literary variety especially in morphology and less in the lexicon and pronunciation. Other differences are marginal.

Writing system

In writing, initially, the medieval Latin alphabet was used and for sounds not present in Latin, digraphs were used. In the early 15th century, the religious reformer Jan Hus replaced the digraphs by single letters with diacritics ("háček" for the palatal/palatalised consonants – ť, ď, ň, ř, š, ť, ž; "čárka" and for long vowels – á, é, í, ó, ú, ý). The only digraph surviving in modern Czech is ch. Long u might have a ring ů, coming from the chain of changes ó>uo>ů.

Typology

Czech along with Slovak, Polish, and the Upper and Low Sorbian belongs to the western Slavonic group of languages. However, Czech separated from the other Slavonic languages by a number of changes, most of which took place in the 10th through 16th centuries (sound changes such as a' > ě, g > h, r'> ř). In the 15th century, Czech lost the dual number and two of the Slavic past tenses – the aorist and imperfect. The verbal aspect had grown more significant and the number of declensions had increased.

2.2 Czech in the Digital Sphere

The Czech republic has .cz as the top level Internet domain. It came into effect in January 1993 after the split of the former Czechoslovakia, which had the domain .cs. As of 21 October 2021, 1,412,102 websites with the top level domain .cz were registered. There were 9.43 million internet users in Czechia in January 2021.[2] The number of internet users in Czechia increased by 123 thousand (+1.3%) between 2020 and 2021. Internet penetration in Czechia stood at 88.0% in January 2021.

There were 7.39 million social media users in Czechia in January 2021. The number of social media users in Czechia increased by 480 thousand (+7.0%) between 2020 and 2021. The number of social media users in Czechia was equivalent to 69% of the total population in January 2021.

[2] According to https://datareportal.com/reports/digital-2021-czechia

3 What is Language Technology?

Natural language[3] 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.

[3] 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).Popular applications 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.[4]

4 Language Technology for Czech

There is quite a large number of data and LT tools for Czech, mostly available from the repository Lindat maintained by the research infrastructure LINDAT/CLARIAH-CZ.

[4] 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).

4.1 Language Data

Monolingual text corpora

The main source of contemporary Czech data are the corpora of the series SYN (Hnátková et al., 2014). SYN2000, SYN2005, SYN2010, SYN2015 and SYN2020 are balanced (representative) corpora of written Czech, morphologically annotated, with the approximately same size of 100 million tokens each. Starting with version 2015, it is possible to make private subcorpora according to specific genre or type of the text, or even more detailed.

SYN2006PUB, SYN2009PUB and SYN2013PUB are corpora of contemporary Czech newspapers and magazines sized 300 MW, 700 MW and 935 MW, respectively. All of the SYN corpora are joined into the single corpus, the last version being SYN v9 (Křen et al., 2021), called the *corpus of contemporary written (printed) Czech*. It contains 4.7 GW.

Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0) is a richly annotated and genre-diversified language resource (Hajič et al., 2020). It is a consolidated release of the existing PDT-corpora of Czech data, uniformly annotated using the standard PDT scheme.

There are also several smaller thematically oriented corpora: The Czech Legal Text Treebank (Kríž et al., 2015) is a collection of 1133 manually annotated dependency trees from the judical domain. The medical domain is covered by the multilingual collections from the Khresmoi project (see the following section). For sentiment analysis, there is a large human-annotated Czech social media corpus (Habernal and Brychcín, 2013).

Bi- and multilingual text corpora

Bilingual data is represented mainly by Czech-English corpora. The 4th release of a sentence-parallel Czech-English corpus CzEng 1.0 (Bojar et al., 2011) contains 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation.

The Universal Dependencies project (Zeman and et al., 2021) releases regularly treebanks in many languages mutually aligned on the sentence level. They contain usually about 2 millions of sentences and are automatically annotated on the morphological and syntactic levels.

There are also many small and medium sized multilingual corpora created for various languages within various domains. The most frequent is the medical domain, for instance the data from the Khresmoi project (Aswani et al., 2013).

Audio corpora

Spoken Czech is recorded in the corpora ORAL2013 (2.8 MW) and ORTOFON v1 (more than 1 MW) (Benešová et al., 2016; Kopřivová et al., 2017).

The most important collection of audio data is the multilingual corpus Czech Malach Cross-lingual Speech Retrieval Test Collection (Galuščáková et al., 2017). The package contains Czech recordings of the Visual History Archive which consists of the interviews with the Holocaust survivors. The archive consists of audio recordings, four types of automatic transcripts, manual annotations of selected topics and interviews' metadata. The whole archive contains 353 recordings and 592 hours of interviews.

The corpus Czech Parliament Meetings (Pražák and Šmídl, 2012) contains recordings from the Chamber of Deputies of the Parliament of the Czech Republic. It consists of 88 hours of speech data, which corresponds roughly to 0.5 million tokens.

Video corpus Czech Television News Broadcasts contains not only video data, but also JSON files with annotations of faces that appear in the broadcasts (Hrúz, 2017).

ESIC (Europarl Simultaneous Interpreting Corpus) (Macháček et al., 2021) is a corpus of 370 speeches (10 hours) in English, with manual transcripts, transcribed simultaneous interpreting into Czech and German, and parallel translations. The corpus contains source English videos and audios.

Lexicons

The basic lexicon is MorfFlex (Hajič et al., 2020), the morphological dictionary of Czech, with full inflectional information for every wordform, which is coded in a positional tag. Wordforms are organised into paradigms according to their formal morphological behavior. The paradigm (set of wordforms) is identified by a unique lemma.

DeriNet (Vidra et al., 2021) is a lexical network which models derivational relations in the lexicon of Czech. Nodes of the network correspond to Czech lexemes, while edges represent word-formational relations between a derived word and its base word/words. The latest release covers more than 1 MW.

Valency dictionary VALLEX provides information on the valency structure (combinatorial potential) of verbs in their particular senses. The latest version VALLEX 4.0 (Lopatková et al., 2020) describes almost 4,700 Czech verbs in more than 11,000 lexical units, i. e. given verbs in the given senses. Similar structure has the lexicon PDT-Vallex (Urešová et al., 2021a): Czech Valency lexicon linked to Prague Dependency treebanks. There is also NomVallex (Kolářová et al., 2020) – the lexicon describing valency of Czech deverbal nouns. In order to facilitate comparison, it also contains abbreviated entries of the source verbs of these nouns from the Vallex lexicon and simplified entries of the covered nouns from the PDT-Vallex lexicon.

The SynSemClass 3.5 (Urešová et al., 2021b) synonym verb lexicon investigates semantic equivalence of verb senses and their valency behavior in parallel Czech-English and German-English language resources, i. e., it relates verb meanings with respect to contextually-based verb synonymy.

Models and grammars

Neural Monkey toolkit (Libovický et al., 2020) for Czech and English is multipurpose. It solves four NLP tasks: machine translation, image captioning, sentiment analysis, and summarisation.

NameTag (Straková and Straka, 2020) is name entity recognition tool for English, German, Dutch, Spanish and Czech. NameTag 2 recognises nested entities (embedded entities) of arbitrary depth.

A model for sentiment analysis (Vysušilová and Straka, 2021) uses the Czech version of BERT model, RobeCzech.

4.2 Language Technologies and Tools

Text analysis

UDPipe (Straka, 2020) is a trainable pipeline for segmentation, tokenisation, POS tagging, morphological analysis, lemmatisation and dependency parsing of raw texts.

MorphoDiTa: Morphological Dictionary and Tagger (Straka and Straková, 2015) is an open-source tool for morphological analysis of texts. It performs morphological analysis, morphological generation, tagging and tokenisation and is distributed as a standalone tool or a library, along with trained linguistic models.

Korektor (Straka and Richter, 2015) is a statistical spellchecker and grammar checker.

Parsito (Straka, 2015) is a fast open-source dependency parser. It has very high accuracy and achieves a throughput of 30,000 words per second. Parsito can be trained on any input data without feature engineering, because it utilises an artificial neural network classifier. Trained models for all treebanks from the Universal Dependencies project are available.

Speech Processing

There is series of Voice Reader tools that provide audio-text processing. Voice Reader 15 converts any kind of text into audio files. The conversion is available in up to 45 languages depending on the version. Voice Reader Home[5] is a text-to-speech synthesis tool. It generates audio files from any (written) text, e. g. emails, e-pub or PDF files, which can be read aloud on any mobile device and desktop PC. Voice Reader Web is a web-based service to read out loud web content in order to make it accessible for hearing impaired people.

Media Studio[6] is a platform for the creation of subtitles from screenplay and media files as well as for translating subtitle content using a custom subtitle machine translation workflow.

The project ELITR[7] (Bojar et al., 2020) is able to recognise and transcript audio input from about 40 languages and translate them online into another language.

Machine translation

The best performing tool for Czech – English translation is the deep-learning system CUBBITT (Popel et al., 2021). In a context-aware blind evaluation by human judges, CUBBITT significantly outperformed professional-agency English-to-Czech news translation in preserving text meaning (translation adequacy). While human translation is still rated as more fluent, CUBBITT is shown to be substantially more fluent than previous state-of-the-art systems. Moreover, most participants of a Translation Turing test struggle to distinguish CUBBITT translations from human translations.

Information Extraction and Information Retrieval

Information extraction from EIA (Environmental impact assessment) documents (Lukšová and Hladká, 2015) is a rule-based extraction system to mine Czech EIA documents.

KER (Libovický, 2016) is a keyword extractor that was designed for scanned texts in Czech and English.

The aforementioned project ELITR also aims to make automatic minuting and summarisation work.

4.3 Projects, Initiatives, Stakeholders

There is a document referred to as The National Artificial Intelligence Strategy of the Czech Republic[8] which was released in 2019 by the Ministry of Industry and Trade of the Czech republic, in which the national AI strategy is presented for the years 2019 – 2030. NLP is mentioned there among disciplines related to human-machine interaction as one of the prominent fields to be supported. At the same time, AICzechia,[9] a national initiative for cooperation between Czech workplaces and teams operating in the field of artificial intelligence, was

[5] https://www.linguatec.de/en/text-to-speech/voice-reader-home-15

[6] https://omniscien.com/media-studio

[7] https://elitr.eu

[8] https://www.mpo.cz/assets/en/guidepost/for-the-media/press-releases/2019/5/NAIS_eng_web.pdf

[9] https://www.aiczechia.cz

established. In the field of NLP applications, it wants to target traditional areas such as defence/security, media and government, but also new domains such as social networks, smart homes or business support. It will maintain and expand activities in international organisations in the field (META-NET Presidency, CLARIN ERIC Committees, LT Innovate Membership, BDVA, ISCA, ACL, IEEE, ELRA and LDC).

The main infrastructure in the field of NLP is LINDAT/CLARIAH-CZ. It is a unique research infrastructure, which deals primarily with language data but also with other digital resources and tools for their exploitation, maintenance and enhancement and offers them to research community, to industry for the development of applications and also directly to the public domain.

Selection of the most important projects in the last 5 years

The most important project in the field of NLP is undoubtedly LINDAT/CLARIAH-CZ – Research Infrastructure for Language Technologies. It brings together all the achievements at a unique place which makes it easily accessible to the wide public.

Universal Dependencies is a project that seeks to develop cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on (universal) Stanford dependencies, Google universal part-of-speech tags, and the Interset interlingua for morphosyntactic tagsets. It publishes 2 versions every year, the latest being (Zeman and et al., 2021).

The outstanding results were achieved in the project Laryngo Voice (Matoušek et al., 2019) – automatic voice conservation and reconstruction with focus on patients after a total laryngectomy. The corpus built in this project contains Czech speech of laryngectomy patients recorded before a surgery causing their voice to be lost in order to preserve the voice which can be later used for a personalised text-to-speech system. Individual utterances were selected from the language by a special algorithm to cover as many phonetic and prosodic features as possible.

In the field of automatic online transcription and translation, the project ELITR, European Live Translator, is being developed. It offers an automatic subtitling system of live meetings and conference presentations and provides a system of spoken language translation (interpreting). In the future, the project aims to design and implement automatic minuting, i. e., create structured summaries from automatic transcripts of discussions and a summary of spoken speeches.

In general, there are several outstanding teams in Czech universities working on all subfields of NLP. They are especially Charles University in Prague, University of West Bohemia, Czech Technical University, Technical University of Liberec, Masaryk University in Brno, Brno University of Technology and Palacký University in Olomouc.

Apart from academia, NLP is being carried on in many private companies, usually (but not always) with a narrower focus.

Results and Applications of the last 5 years

In the field of machine translation, several systems were handed over for public usage. The most successful are CUBBITT and UDPipe, already mentioned.

The new edition of the PDT treebank, namely PDT-C, was also already mentioned.

The work on speech recognition and indexation for digitised oral history archives MALACH (Holocaust survivors' testimony, archive of the Institute for the Study of Totalitarian Regimes)[10] continues and new tools are being developed.

[10] https://ufal.mff.cuni.cz/malach/en

The Alquist Dialogue System[11] is the social bot developed by a team of students from the Czech Technical University in Prague. Alquist is an advanced Conversational AI bot carrying an entertaining and engaging conversation with humans on popular topics, such as movies, sports, news, etc. In 2017 and 2018, it gained the second place in the Alexa Prize contests in competition with over 100 academic teams from around the world.

5 Cross-Language Comparison

The LT field[12] 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[13] 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)
  - 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.)

[11] http://alquistai.com

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

[13] 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.

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 $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[14]

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[15] 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 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.[16]

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

[14] 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.

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

[16] 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.

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 1 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,[17] 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, 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

[17] In addition to the languages listed in Table 1, 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 1: 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)

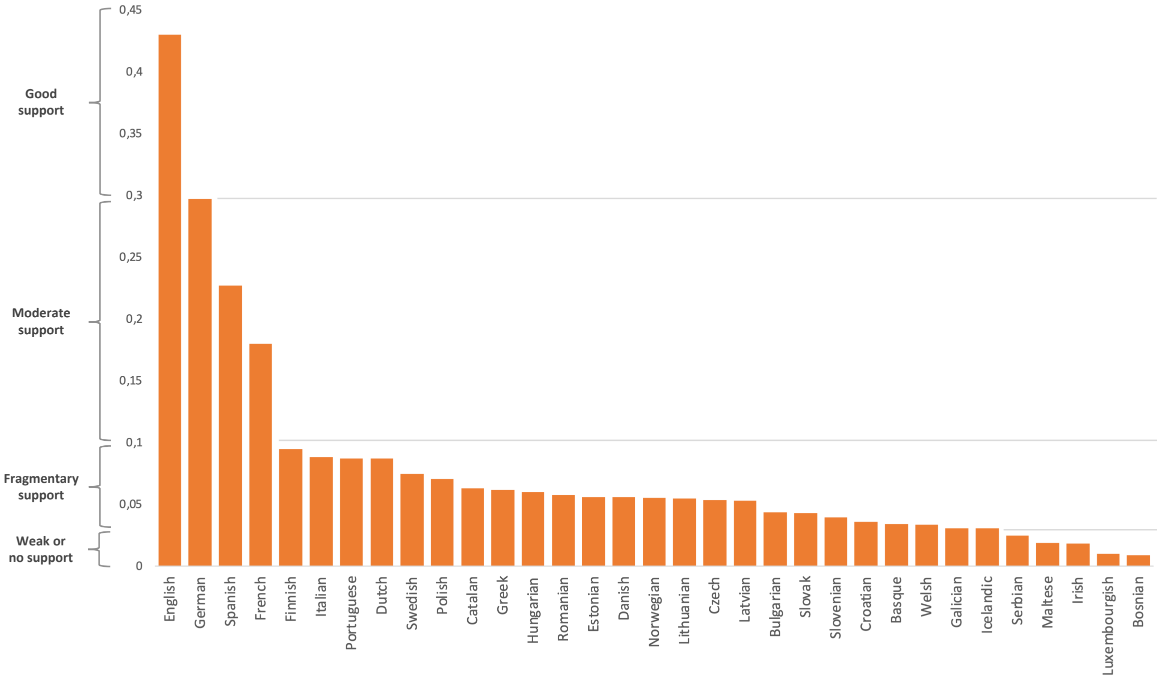

Figure 1: Overall state of technology support for selected European languages (2022)

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

Natural language processing has a privileged position among AI disciplines in the Czech Republic in terms of the quantity and quality of publications, the quantity and prestige of international projects (including US agencies NSF, ONR, DARPA and IARPA) and industrial collaborations or firms founded or working closely with research groups (Phonexia, Lexical Computing, Lingea, SpeechTech, MemSource, Newton Technologies, Newton Media, Replaywell and others). Many of them are active abroad and realise most of their turnover outside the Czech Republic. The field already contributes several hundred jobs to the economy (labs and bonded firms alone).

The future of technological development lies in connecting modalities (speech, text, video), developing algorithms on inaccurately described or completely unscripted data available in large quantities on the internet, improving robustness (e. g. when processing data from a new type of phone or in a new dialect) and accelerating the development cycle through end-to-end training.

In the area of applications, the traditional areas such as defense/security, media and government should be promoted, but there are also new targets such as social networks, smart homes or linking speech and text with business process support.

In the field of design and organisation, it is important to maintain and increase excellence in European, American and national projects, to continue to work with the international scientific community, including the continuation of the already started internationalisation of our teams, and to work on a stronger connection between *speech* and *text* communities. It is reasonable to maintain and expand activities in international organisations in the field (META-NET presidency, CLARIN ERIC committees, membership in LT Innovate, BDVA, ISCA, ACL, IEEE, ELRA and LDC).

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.

Niraj Aswani, Thomas Beckers, Erich Birngruber, Célia Boyer, Andreas Burner, Jakub Bystroň, Khalid Choukri, Sarah Cruchet, Hamish Cunningham, Jan Dědek, Ljiljana Dolamic, René Donner, Sebastian Dungs, Ivan Eggel, Antonio Foncubierta, Norbert Fuhr, Adam Funk, Alba Herrera, Arnaud Gaudinat, Georgi Georgiev, Julien Gobeill, Lorraine Goeuriot, Paz Gomez, Mark Greenwood, Manfred Gschwandtner, Allan Hanbury, Jan Hajič, Jaroslava Hlaváčová, Markus Holzer, Gareth Jones, Blanca Jordán, Matthias Jordan, Klemens Kaderk, Franz Kainberger, Liadh Kelly, Sascha Kriewel, Marlene Kritz, Georg Langs, Nolan Lawson, Dimitrios Markonis, Iván Martínez, Vassil Momtchev, Alexandre Masselot, Hélène Mazo, Henning Müller, Pavel Pecina, Konstantin Pentchev, Deyan Peychev, Natalia Pletneva, Diana Pottecher, Angus Roberts, Patrick Ruch, Matthias Samwald, Priscille Schneller, Veronika Stefanov, Miguel Tinte, Zdeňka Urešová, Alejandro Vargas, and Dina Vishnyakova. Khresmoi - multilingual semantic search of medical text and images. In Christoph Lehmann, Elske Ammenwerth, and Christian Nøhr, editors, *MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics*, volume 192 of *Studies in Health Technology and Informatics*, pages 1266–1266, Amsterdam, Netherlands, 2013. International Medical Informatics Association, IOS Press. ISBN 978-1-61499-288-2.

Lucie Benešová, Michal Křen, and Martina Waclawičová. ORAL2013: balanced corpus of informal spoken czech (transcriptions & audio), 2016. URL http://hdl.handle.net/11234/1-1848. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. Czech-english parallel corpus 1.0 (CzEng 1.0), 2011. URL http://hdl.handle.net/11234/1-1458. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Ebrahim Ansari, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. ELITR: European live translator. In *Proceedings of the 22st Annual Conference of the European Association for Machine Translation (2020)*, pages 463–464, Lisboa, Portugal, 2020. Universitat d'Alacant, European Association for Machine Translation. ISBN 978-989-33-0589-8.

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

Petra Galuščáková, Pavel Pecina, Petra Hoffmannová, Jan Hajič, Pavel Ircing, and Jan Švec. Czech malach cross-lingual speech retrieval test collection, 2017. URL http://hdl.handle.net/11234/1-1912. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ivan Habernal and Tomáš Brychcín. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of RANLP 2013*, page TBD. Association for Computational Linguistics, 2013. URL TBD.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. Prague dependency treebank - consolidated 1.0 (PDT-c 1.0), 2020.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. MorfFlex CZ 2.0, 2020. URL http://hdl.handle.net/11234/1-3186. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

M. Hnátková, M. Křen, P. Procházka, and H. Skoumalová. The syn-series corpora of written czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 160—164, Reykjavík, Island, 2014.

Marek Hrúz. Czech television news broadcasting faces, 2017. URL http://hdl.handle.net/11234/1-2545. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Veronika Kolářová, Anna Vernerová, and Jana Klímová. NomVallex i., 2020. URL http://hdl.handle.net/11234/1-3420. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Marie Kopřivová, Zuzana Komrsková, David Lukeš, Petra Poukarová, and Marie Škarpová. ORTOFON v1: balanced corpus of informal spoken czech with multi-tier transcription (transcriptions & audio), 2017. URL http://hdl.handle.net/11234/1-2579. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. SYN v9: large corpus of written czech, 2021. URL http://hdl.handle.net/11234/1-4635. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vincent Kríž, Barbora Hladká, and Zdeňka Urešová. Czech legal text treebank, 2015. URL http://hdl.handle.net/11234/1-1516. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jindřich Libovický. KER - keyword extractor, 2016. URL http://hdl.handle.net/11234/1-1650. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jindřich Libovický, Rudolf Rosa, Jindřich Helcl, and Martin Popel. Czech image captioning, machine translation, sentiment analysis and summarization (neural monkey models), 2020. URL http://hdl.handle.net/11234/1-3145. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Markéta Lopatková, Václava Kettnerová, Anna Vernerová, Eduard Bejček, and Zdeněk Žabokrtský. VALLEX 4.0 (2021-02-12), 2020. URL http://hdl.handle.net/11234/1-3524. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ivana Lukšová and Barbora Hladká. Information extraction from EIA documents, 2015. URL http://hdl.handle.net/11234/1-1515. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. ESIC 1.0 – europarl simultaneous interpreting corpus, 2021. URL http://hdl.handle.net/11234/1-3719. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jindřich Matoušek, Daniel Tihelka, Markéta Jůzová, Martin Grůber, Jakub Vít, and Barbora Řepová. Database of speech corpora of czech laryngectomy patients, 2019. URL http://hdl.handle.net/11234/1-3142. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Popel, Markéta Tomková, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. CUBBITT translation models (en-cs) (v1.0), 2021. URL http://hdl.handle.net/11234/1-3733. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Aleš Pražák and Luboš Šmídl. Czech parliament meetings, 2012. URL http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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 Straka. Parsito, 2015. URL http://hdl.handle.net/11234/1-1584. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Milan Straka. UDPipe 2, 2020.

Milan Straka and Michal Richter. Korektor 2, 2015. URL http://hdl.handle.net/11234/1-1469. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Milan Straka and Jana Straková. Morphodita, 2015.

Jana Straková and Milan Straka. NameTag 2 models (2021-09-16), 2020. URL http://hdl.handle.net/11234/1-3773. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

Zdeňka Urešová, Alevtina Bémová, Eva Fučíková, Jan Hajič, Veronika Kolářová, Marie Mikulová, Petr Pajas, Jarmila Panevová, and Jan Štěpánek. PDT-vallex: Czech valency lexicon linked to treebanks 4.0 (PDT-vallex 4.0), 2021a. URL http://hdl.handle.net/11234/1-3499. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, Jan Hajič, Karolina Zaczynska, and Georg Rehm. SynSemClass 3.5, 2021b. URL http://hdl.handle.net/11234/1-3750. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. DeriNet 2.1, 2021. URL http://hdl.handle.net/11234/1-3765. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Petra Vysušilová and Milan Straka. Sentiment analysis (czech model), 2021. URL http://hdl.handle.net/11234/1-4601. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman and et al. Universal dependencies 2.9, 2021. URL http://hdl.handle.net/11234/1-4611. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.